PAC-BAYESIAN BOUNDS FOR RANDOMIZED EMPIRICAL RISK MINIMIZERS

PIERRE ALQUIER

Abstract. The aim of this paper is to generalize the PAC-Bayesian theorems proved by Catoni [6, 8] in the classification setting to more general problems of statistical inference. We show how to control the deviations of the risk of randomized estimators. A particular attention is paid to randomized estimators drawn in a small neighborhood of classical estimators, whose study leads to control the risk of the latter. These results allow to bound the risk of very general estimation procedures, as well as to perform model selection.

Contents

Date: January 12, 2009.

2000 *Mathematics Subject Classification.* Primary 62G08; Secondary 62H30, 68T05, 68T10.

Key words and phrases. Regression estimation, classification, adaptative inference, statistical learning, randomized estimator, empirical risk minimization, empirical bound.

I Would like to thank Professor Olivier Catoni for his kind help and his useful remarks.

1. Introduction

The aim of this paper is to perform statistical inference with observations in a possibly large dimensional space. Let us first introduce the notations.

1.1. **General notations.** Let $N \in \mathbb{N}^*$ be the number of observations. Let $(\mathcal{Z}, \mathcal{B})$ be a measurable space and P_1, ..., P_N be N probability measures on this space, unknown to the statistician. We assume that

$$(Z_1, ..., Z_N)$$

is the canonical process on

$$\left(\mathcal{Z}^N, \mathcal{B}^{\otimes N}, P_1 \otimes ... \otimes P_N\right).$$

Definition 1.1. Let us put

$$\mathbb{P} = P_1 \otimes ... \otimes P_N,$$

and

$$\overline{\mathbb{P}} = \frac{1}{N} \sum_{i=1}^{N} \delta_{Z_i}.$$

We want to perform statistical inference on a general parameter space Θ, with respect to some loss function

$$\ell_\theta : \mathcal{Z} \to \mathbb{R}, \qquad \theta \in \Theta.$$

Definition 1.2 (Risk functions)**.** We introduce, for any $\theta \in \Theta$,

$$r(\theta) = \overline{\mathbb{P}}\left(\ell_\theta\right) = \frac{1}{N} \sum_{i=1}^{N} \ell_\theta\left(Z_i\right),$$

the empirical risk function, and

$$R(\theta) = \mathbb{P}(\ell_\theta) = \frac{1}{N} \sum_{i=1}^{N} P_i\left(\ell_\theta\right),$$

the risk function.

We now describe three classical problems in statistics that fit the general context described above.

Example 1.1 (Classification)**.** We assume that $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is a set of objects and $\mathcal{Y}$ a finite set of possible labels for these objects. Consider a set of classification functions $\{f_\theta : \mathcal{X} \to \mathcal{Y}, \theta \in \Theta\}$ which assign to each object a label. Let us put, for any $z = (x, y) \in \mathbb{Z}$, $\ell_\theta(z) = \psi\left(f_\theta(x), y\right)$ where ψ is some symmetric discrepancy measure. The most usual case is to use the 0-1 loss function $\psi(y, y') = \delta_y(y')$. If moreover $|\mathcal{Y}| = 2$ we can decide that $\mathcal{Y} = \{-1, +1\}$ and set $\psi(y, y') = \mathbb{1}_{\mathbb{R}_+^*}(yy')$. However, in many practical situations, algorithmic considerations lead to use a convex upper bound of this loss function, like

$$\psi(y, y') = (1 - yy')_+ = \max(1 - yy', 0), \quad \text{the "hinge loss"},$$

$$\psi(y, y') = \exp(-yy'), \quad \text{the exponential loss},$$

$$\psi(y, y') = (1 - yy')^2, \quad \text{the least square loss}.$$

For example, Cortes and Vapnik [10] generalized the SVM technique to non-separable data using the hinge loss, while Schapire, Freund, Bartlett and Lee [19] gave a statistical interpretation of boosting algorithm thanks to the exponential loss. See Zhang [22] for a complete study of the performance of classification methods using

these loss functions. Remark that in this case, f_θ is allowed to take any real value, and not only -1 or $+1$, although the labels Y_i in the training set are either -1 or $+1$.

Example 1.2 (Regression estimation)**.** The context is the same except that the label set $\mathcal{Y}$ is infinite, in most case it is $\mathbb{R}$ or an interval of $\mathbb{R}$. Here, the most usual case is the regression with quadratic loss, with $\psi(y,y') = (y-y')^2$, however, more general cases can be studied like the l^p loss $\psi(y,y') = (y-y')^p$ for some $p \geq 1$.

Example 1.3 (Density estimation)**.** Here, we assume that $P_1 = ... = P_N = P$ and consequently that $\mathbb{P} = P^{\otimes N}$, and we want to estimate the density $f = dP/d\mu$ of P with respect to a known measure μ. We assume that we are given a set of probability measures $\{Q_\theta, \theta \in \Theta\}$ with densities $q_\theta = dQ_\theta/d\mu$ and we use the loss function $\ell_\theta(z) = -\log\left[q_\theta(z)\right]$. Indeed in this case, we can write under suitable hypotheses

$$R(\theta) = P\left(-\log \circ q_\theta\right) = P\left(-\log \circ \frac{dQ_\theta}{d\mu}\right) = P\left(\log \circ \frac{dP}{dQ_\theta}\right) + P\left(\log \circ \frac{d\mu}{dP}\right)$$
$$= \mathcal{K}\left(P, Q_\theta\right) - P\left(\log \circ f\right),$$

showing that the risk is the Kullback-Leibler divergence between P and Q_θ up to a constant (the definition of $\mathcal{K}$ is reminded in this paper, see Definition 1.8 page 5).

In each case the objective is to estimate $\arg\min R$ on the basis of the observations $Z_1, ..., Z_N$ - presumably using in some way or another the value of the empirical risk. We have to notice that when the space Θ is large or complex (for example a vector space with large dimension), $\arg\min R$ and $\arg\min r$ can be very different. This does not happen if Θ is simple (for example a vector space with small dimension), but such a case is less interesting as we have to eliminate a lot of dimensions in Θ before proceeding to statistical inference with no guarantees that these directions are not relevant.

1.2. **Statistical learning theory and PAC-Bayesian point of view.** The learning theory point of view introduced by Vapnik and Cervonenkis ([9], see Vapnik [21] for a presentation of the main results in English) gives a setting that proved to be adapted to deal with estimation problems in large dimension. This point of view received an important interest over the past few years, see for example the well-known books of Devroye, Gyï£¡rfi and Lugosi [11], Friedman, Hastie and Tibshirani [12] or more recently the paper by Boucheron, Bousquet and Lugosi [5] and the references therein, for a state of the art.

The idea of Vapnik and Cervonenkis is to introduce a structure, namely a family of submodels Θ_1, Θ_2, ... The problem of model selection then arises: we must choose the submodel Θ_k in which the minimization of the empirical risk r will lead to the smallest possible value for the real risk R. This choice requires to estimate the complexity of submodels Θ_k. An example of complexity measure is the so-called Vapnik Cervonenkis dimension or VC-dimension, see [9, 21].

The PAC-Bayesian point of view, introduced in the context of classification by McAllester [16, 17] is based on the following remark: while classical measures of complexity (like VC-dimension) require theoretical results on the submodels, the introduction of a probability measure π on the model Θ allows to measure empirically the complexity of every submodel. In a more technical point of view, we will see later that π allows a generalization of the so-called union bound (see [17] for example). This point of view might be compared with Rissanen's work on MDL (Minimum Description Length, see [18]) making a link between statistical inference and information theory, and $-\log\pi(\theta)$ can be seen as the length of a code for the parameter θ (at least when Θ is finite).

The PAC-Bayesian point of view was developed in more contexts (classification, least square regression and density estimation) by Catoni [7], and then improved in the context of classification by Catoni [6], Audibert [3] and in the context of least square regression by Audibert [2] and of regression with a general loss in our PhD thesis [1]. The most recent work in the context of classification by Catoni [8] improves the upper-bound given on the risk of the PAC-Bayesian estimators, leading to purely empirical bounds that allow to perform model selection with no assumption on the probability measure $\mathbb{P}$. The aim of this work is to extend these results to the very general context of statistical inference introduced in subsection 1.1, that includes classification, regression with a general loss function and density estimation.

Let us introduce our estimators.

Definition 1.3. Let us assume that we have a family of functions

$$\psi_\theta^i : \mathcal{Z} \to \mathbb{R} \cup \{+\infty\}$$

indexed by i in a finite or countable set I and by $\theta \in \Theta$. For every $i \in I$ we choose:

$$\hat{\theta}_i \in \arg\min_{\theta\in\Theta} \overline{\mathbb{P}}\left(\psi_\theta^i\right).$$

Example 1.4 (Empirical risk minimization and model selection)**.** If we take $I = \{0\}$ we can choose $\psi_\theta^0(z) = l_\theta(z)$ and we obtain $\overline{\mathbb{P}}\left(\psi_\theta^0\right) = r(\theta)$ and so

$$\hat{\theta}^0 = \arg\min_{\theta\in\Theta} r(\theta)$$

the empirical risk minimizer. In the case where the dimension of Θ is large, we can choose several submodels indexed by a finite or countable family I: $(\Theta_i, i \in I)$. In order to obtain

$$\hat{\theta}^i = \arg\min_{\theta\in\Theta_i} r(\theta)$$

we can put

$$\psi_\theta^i(.) = \begin{cases} l_\theta(.) & \text{if} \quad \theta \in \Theta_i \\ \\ +\infty & \text{otherwise.} \end{cases}$$

The problem of the selection of the $\hat{\theta}_i$ with the smallest possible risk (so-called model selection problem) can be solved with the help of PAC-Bayesian bounds.

Note that PAC-Bayesian bounds given by Catoni [6, 7, 8] usually apply to "randomized estimators". More formally, let us introduce a σ-algebra $\mathcal{T}$ on Θ and a probability measure π on the measurable space $(\Theta, \mathcal{T})$. We will need the following definitions.

Definition 1.4. For any measurable set $(E, \mathcal{E})$, we let $\mathcal{M}_+^1(E)$ denote the set of all probability measures on the measurable space $(E, \mathcal{E})$.

Definition 1.5. In order to generalize the notion of estimator (a measurable function $\mathcal{Z}^N \to \Theta$), we call a randomized estimator any function $\rho : \mathcal{Z}^N \to \mathcal{M}_+^1(\Theta)$ that is a regular conditional probability measure. For the sake of simplicity, the sample being given, we will write ρ instead of $\rho(Z_1, ..., Z_N)$.

PAC-Bayesian bounds for randomized estimators are usually given for their mean risk

$$\int_{\theta\in\Theta} R(\theta) d\rho(\theta),$$

whereas here we will rather focus on $R(\tilde{\theta})$, where $\tilde{\theta}$ is drawn from ρ and ρ is highly concentrated around a "classical" (deterministic) estimator $\hat{\theta}_i$.

1.3. **Truncation of the risk.** In this subsection, we introduce a truncated version of the relative risk of two parameters θ and θ'.

Definition 1.6. We put, for any $\lambda \in \mathbb{R}_+^*$ and $(\theta, \theta') \in \Theta^2$

$$R_\lambda(\theta, \theta') = \mathbb{P}\left[(\ell_\theta - \ell_{\theta'}) \wedge \frac{N}{\lambda}\right].$$

Note of course that if $\mathbb{P}$-almost surely, we have $\ell_\theta - \ell_{\theta'} \leq N/\lambda$ then $R_\lambda(\theta, \theta') = R(\theta) - R(\theta')$.

In what follows, we will give empirical bounds on $R_\lambda(\theta, \theta')$ for some θ and θ' chosen by some statistical procedure. One can wonder why we prefer to bound this truncated version of the risk instead of $R(\theta) - R(\theta')$. The reason is the following. In this paper, we want to give bounds that hold with no particular assumption on the unknown data distribution $\mathbb{P}$. However, it is clear that we cannot obtain a purely empirical bound on $R(\theta) - R(\theta')$ with no assumption on the data distribution, as it is shown by the following example.

Example 1.5. Let us choose $c > 0$ and $\lambda > 0$. We assume that $P_1 = ... = P_N$ and that $\Theta = \{\theta, \theta'\}$ with $l_{\theta'}(z) = 0$. We put $l_\theta(Z) = cN$ with probability $1/N$ and 0 otherwise. Then we have $R(\theta') = 0$ and

$$R(\theta) = \frac{1}{N} cN + \left(1 - \frac{1}{N}\right) 0 = c$$

while $r(\theta') = 0$ and with probability at least $(1 - 1/N)^N \simeq \exp(-1)$ we also have $r(\theta) = 0$, this means that we cannot upper bound precisely $R(\theta) - R(\theta')$ by empirical quantities with no assumption.

So, we introduce the truncation of the risk. However, two remarks shall be made. First, in the case of a bounded loss function ℓ, with a large enough ratio N/λ we have $R_\lambda(\theta, \theta') = R(\theta) - R(\theta')$.

In the general case, if we want to upper bound $R(\theta) - R(\theta')$ we can make additional hypotheses on the data distribution, ensuring that we can dispose of a (known) upper-bound :

$$\Delta_\lambda(\theta, \theta') \geq R(\theta) - R(\theta') - R_\lambda(\theta, \theta')$$

as it is done in our PhD Thesis [1]. For the sake of completeness, such an upper bound is given in the Appendix, page 28.

1.4. **Main tools.** In this subsection, we give two lemmas that will be useful in order to build PAC-Bayesian theorems. First, let us recall the following definition. In this whole subsection, we assume that $(E, \mathcal{E})$ is an arbitrary measurable space.

Definition 1.7. For any measurable function $h : E \to \mathbb{R}$, for any measure $m \in \mathcal{M}_+^1(E)$ we put

$$m(h) = \sup_{B \in \mathbb{R}} \int_E [h(x) \wedge B]\, m(dx).$$

Definition 1.8 (Kullback-Leibler divergence)**.** Given a measurable space $(E, \mathcal{E})$, we define , for any $(m, n) \in [\mathcal{M}_+^1(E)]^2$, the Kullback-Leibler divergence function

$$\mathcal{K}(m, n) = \begin{cases} \displaystyle\int_E dm(e) \left\{ \log\left[\frac{dm}{dn}(e)\right]\right\} & \text{if } m \ll n, \\ +\infty & \text{otherwise.} \end{cases}$$

Lemma 1.1 (Legendre transform of the Kullback divergence function)**.** *For any $n \in \mathcal{M}_+^1(E)$, for any measurable function $h : E \to \mathbb{R}$ such that $n(\exp \circ h) < +\infty$ we have*

$$\log n(\exp \circ h) = \sup_{m \in \mathcal{M}_+^1(E)} \Big(m(h) - \mathcal{K}(m,n) \Big), \tag{1.1}$$

where by convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of n, the supremum with respect to m in the right-hand side is reached for the Gibbs distribution, $n_{\exp(h)}$ given by:

$$\forall e \in E, \quad \frac{dn_{\exp(h)}}{dn}(e) = \frac{\exp[h(e)]}{\pi(\exp \circ h)}.$$

The proof of this lemma is given at the end of the paper, in a section devoted to proofs (subsection 5.1 page 15). We now state another lemma that will be useful in the sequel. First, we need the following definition.

Definition 1.9. We put, for any $\alpha \in \mathbb{R}_+^*$,

$$\begin{aligned} \Phi_\alpha : \quad &]-\infty, 1/\alpha[\to \mathbb{R} \\ & \qquad t \mapsto -\frac{\log(1-\alpha t)}{\alpha}. \end{aligned}$$

Note that Φ_α is invertible, that for any $u \in \mathbb{R}$,

$$\Phi_\alpha^{-1}(u) = \frac{1 - \exp(-\alpha u)}{\alpha} \leq u,$$

and that $\frac{2(\Phi_\alpha(x) - x)}{\alpha x^2} \xrightarrow[x \to 0]{} 1$. Also note that for $\alpha > 0$, Φ_α is convex and that $\Phi_\alpha(x) \geq x$. An elementary study of this function also proves that for any $C > 0$, for any $\alpha \in]0, 1/(2C)[$ and any $p \in [0, C]$ we have:

$$\Phi_\alpha(p) \leq p + \frac{\alpha p^2}{2}.$$

We can now give the lemma.

Lemma 1.2. *We have, for any $\lambda \in \mathbb{R}_+^*$, for any $a \in]0,1]$, for any $(\theta, \theta') \in \Theta^2$,*

$$\mathbb{P} \exp \left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R_{\frac{\lambda}{a}}(\theta, \theta') \right] - \frac{\lambda}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} \left[(\ell_\theta - \ell_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right] \right\} = 1.$$

The proof is almost trivial, we give it now in order to emphasize the role of the truncation and of the change of variable.

Proof. For any $\lambda \in \mathbb{R}_+^*$, for any $(\theta, \theta') \in \Theta^2$,

$$\begin{aligned} \mathbb{P} \exp & \left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R_{\frac{\lambda}{a}}(\theta, \theta') \right] - \frac{\lambda}{N} \sum_{i=1}^N \Phi_{\frac{\lambda}{N}} \left[(\ell_\theta - \ell_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right] \right\} \\ &= \mathbb{P} \exp \Bigg\{ \sum_{i=1}^N \Bigg(\log \left[1 - \frac{\lambda}{N} \left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right) \right] \\ &\qquad\qquad - \log \left[1 - \frac{\lambda}{N} P_i \left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right) \right] \Bigg) \Bigg\} \\ &= \mathbb{P} \left[\prod_{i=1}^N \frac{1 - \frac{\lambda}{N} \left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right)}{1 - \frac{\lambda}{N} P_i \left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right)} \right] \end{aligned}$$

$$= \prod_{i=1}^{N} P_i \left[\frac{1 - \frac{\lambda}{N}\left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda}\right)}{1 - \frac{\lambda}{N} P_i \left((l_\theta - l_{\theta'})(Z_i) \wedge \frac{aN}{\lambda}\right)} \right] = 1.$$

□

Note that this lemma will be used as an alternative to Hoeffding's or Bernstein's (see [13, 4]) inequalities in order to prove PAC inequalities.

1.5. **A basic PAC-Bayesian Theorem.** Let us integrate Lemma 1.2 with respect to (θ, θ') with a given probability measure $n = \pi \otimes \pi'$ with $(\pi, \pi') \in [\mathcal{M}_+^1(\Theta)]^2$. Applying Fubini-Tonelli Theorem we obtain:

$$\mathbb{P}\left\{ \int_{(\theta,\theta')\in\Theta^2} d(\pi\otimes\pi')(\theta,\theta') \exp\left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R_{\frac{\lambda}{a}}(\theta,\theta') \right] - \frac{\lambda}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[\left(\ell_\theta - \ell_{\theta'}\right)(Z_i) \wedge \frac{aN}{\lambda} \right] \right\} \right\} = 1. \tag{1.2}$$

This implies that for any $(\rho, \rho') \in [\mathcal{M}_+^1(\Theta)]^2$,

$$\mathbb{P}\left\{ \int_{(\theta,\theta')\in\Theta^2} d(\rho\otimes\rho')(\theta,\theta') \exp\left\{ \lambda \Phi_{\frac{\lambda}{N}} \left[R_{\frac{\lambda}{a}}(\theta,\theta') \right] - \frac{\lambda}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[\left(\ell_\theta - \ell_{\theta'}\right)(Z_i) \wedge \frac{aN}{\lambda} \right] - \log\left[\frac{d(\rho\otimes\rho')}{d(\pi\otimes\pi')}(\theta,\theta') \right] \right\} \right\} \leq 1.$$

(This inequality becomes an equality when $\pi \ll \rho$ and $\pi' \ll \rho'$.)

Theorem 1.3. *Let us assume that we have $(\pi, \pi') \in \mathcal{M}_+^1(\Theta)^2$, and two randomized estimators ρ and ρ'. For any $\varepsilon > 0$, for any $(a, \lambda) \in]0, 1] \times \mathbb{R}_+^*$, with $\mathbb{P}(\rho \otimes \rho')$-probability at least $1 - \varepsilon$ over the sample $(Z_i)_{i=1,\ldots,N}$ and the parameters $(\tilde{\theta}, \tilde{\theta}')$, we have:*

$$R_{\frac{\lambda}{a}}\left(\tilde{\theta}, \tilde{\theta}'\right) \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ \frac{1}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[\left(\ell_{\tilde{\theta}} - \ell_{\tilde{\theta}'}\right)(Z_i) \wedge \frac{aN}{\lambda} \right] + \frac{\log\left[\frac{d\rho}{d\pi}\left(\tilde{\theta}\right)\right] + \log\left[\frac{d\rho'}{d\pi'}\left(\tilde{\theta}'\right)\right] + \log\frac{1}{\varepsilon}}{\lambda} \right\}.$$

In order to provide an interpretation of Theorem 1.3, let us give the following corollary in the bounded case, which is obtained using basic properties of the function Φ given just after Definition 1.9 page 6. In this case, the parameter a is just set to 1.

Corollary 1.4. *Let us assume that for any $(\theta, z) \in \Theta \times \mathcal{Z}, 0 < l_\theta(z) < C$. Let us assume that we have $(\pi, \pi') \in \mathcal{M}_+^1(\Theta)^2$, and two randomized estimators ρ and ρ'. For any $\varepsilon > 0$, for any $\lambda \in]0, N/(2C)]$, with $\mathbb{P}(\rho \otimes \rho')$-probability at least $1 - \varepsilon$ we have:*

$$R(\tilde{\theta}) - R(\tilde{\theta}') \leq \Phi_{\frac{\lambda}{N}}^{-1} \left\{ r(\tilde{\theta}) - r(\tilde{\theta}') + \frac{\lambda}{2N} \overline{\mathbb{P}} \left[\left(l_{\tilde{\theta}} - l_{\tilde{\theta}'}\right)^2 \right] + \frac{\log\left[\frac{d\rho}{d\pi}(\tilde{\theta})\right] + \log\left[\frac{d\rho'}{d\pi'}(\tilde{\theta}')\right] + \log\frac{1}{\varepsilon}}{\lambda} \right\}.$$

We can see that the difference of the "true" risk of the randomized estimators $\tilde{\theta}$ and $\tilde{\theta}'$, drawn independently from ρ and ρ', is upper bounded by the difference of the empirical risk, plus a variance term and a complexity term expressed in terms of the log of the density of the randomized estimator with respect to a given prior. So Theorem 1.3 provides an empirical way to compare the theoretical performance of two randomized estimators, leading to applications in model selection. This paper is devoted to improvements of Theorem 1.3 (we will see in the sequel that this theorem does not necessarily lead to optimal estimators) and to the effective construction of estimators using variants of Theorem 1.3.

Now, note that the choice of the randomized estimators ρ and ρ' is not straightforward. The following theorem, which gives an integrated variant of Theorem 1.3, can be usefull for that purpose.

Theorem 1.5. *Let us assume that we have $(\pi,\pi') \in \mathcal{M}_+^1(\Theta)^2$. For any $\varepsilon > 0$, for any $(a,\lambda) \in]0,1] \times \mathbb{R}_+^*$, with $\mathbb{P}$-probability at least $1-\varepsilon$, for any $(\rho,\rho') \in \mathcal{M}_+^1(\Theta)^2$,*

$$\int_{\Theta^2} R_{\frac{\lambda}{a}}(\theta,\theta') d(\rho\otimes\rho')(\theta,\theta') \\ \leq \Phi_{\frac{\lambda}{N}}^{-1}\Bigg\{\int_{\Theta^2} \frac{1}{N}\sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}}\left[\left(\ell_\theta - \ell_{\theta'}\right)(Z_i) \wedge \frac{aN}{\lambda}\right] d\left(\rho\otimes\rho'\right)(\theta,\theta') \\ + \frac{\mathcal{K}(\rho,\pi) + \mathcal{K}(\rho',\pi') + \log\frac{1}{\varepsilon}}{\lambda}\Bigg\}.$$

The proof is given in subsection 5.2 page 15.

1.6. **Main results of the paper.** In our PhD dissertation [1], a particular case of Theorem 1.5 is given and applied to regression estimation with quadratic loss in a bounded model of finite dimension d. In this particular case, it is shown that the estimators based on the minimization of the right-hand side of Theorem 1.5 do not achieve the optimal rate of convergence: d/N, but only $(d\log N)/N$. A solution is given by Catoni in [7] and consists in replacing the prior π by the so-called "localized prior" $\pi_{\exp(-\beta R)}$ for a given $\beta > 0$. The main problem is that this choice leads to the presence of non-empirical terms in the right-hand side, $\mathcal{K}(\rho,\pi_{\exp(-\beta R)})$.

In Section 2, we give an empirical bound for this term $\mathcal{K}(\rho,\pi_{\exp(-\beta R)})$. We also give a heuristic that leads to this technique of localization.

In Section 3, we show how this result, combined with Theorem 1.5, leads to the effective construction of an estimator that can reach optimal rates of convergence.

The proofs of the theorems stated in this paper are gathered in Section 5.

2. Empirical bound for the localized complexity and localized PAC-Bayesian theorems

2.1. **Mutual information between the sample and the parameter.** Let us consider Theorem 1.5 with $\rho' = \pi' = \delta_{\theta'}$ for a given parameter θ'. For the sake of simplicity, let us assume in this subsection that we are in the bounded case (l_θ bounded by C). Theorem 1.5 ensures that, for any $\lambda \in]0, N/(2C)[$, with $\mathbb{P}$-probability at least $1-\varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,

$$\rho(R) - R(\theta') \leq \rho(r) - r(\theta') + \frac{\lambda}{2N}\overline{\mathbb{P}}\left[\int_\Theta \left(l_\theta - l_{\theta'}\right)^2 d\rho(\theta)\right] + \frac{\mathcal{K}(\rho,\pi) + \log\frac{1}{\varepsilon}}{\lambda}.$$

This is an incitation to choose

$$\rho = \arg\min_{\mu\in\mathcal{M}_+^1(\Theta)} \left[\mu(r) + \frac{\lambda}{2N}\overline{\mathbb{P}}\left[\int_\Theta \left(l_\theta - l_{\theta'}\right)^2 d\rho(\theta)\right] + \frac{\mathcal{K}(\mu,\pi)}{\lambda}\right].$$

However, if we choose to neglect the variance term, we may consider the following randomized estimator:

$$\rho = \arg \min_{\mu \in \mathcal{M}_+^1(\Theta)} \left[\mu(r) + \frac{\mathcal{K}(\mu, \pi)}{\lambda} \right].$$

Actually, in this case, Lemma 1.1 leads to:

$$\rho = \pi_{\exp(-\lambda r)}.$$

Let us remark that, for any $(\rho, \pi) \in \mathcal{M}_+^1(\Theta)$ we have:

$$P\Big[\mathcal{K}(\rho, \pi)\Big] = P\Big[\mathcal{K}(\rho, P(\rho))\Big] + \mathcal{K}(P(\rho), \pi). \tag{2.1}$$

This implies that, for a given data-dependent ρ, the optimal deterministic measure π is $P(\rho)$ in the sense that it minimizes the expectation of $\mathcal{K}(\rho, \pi)$ (left-hand side of Equation 2.1), making it equal to the expectation of $\mathcal{K}(\rho, P(\rho))$. This last quantity is the mutual information between the estimator and the sample.

So, for $\rho = \pi_{\exp(-\lambda r)}$, this is an incitation to replace the prior π with $P\left(\pi_{\exp(-\lambda r)}\right)$. It is then natural to approximate this distribution by $\pi_{\exp(-\lambda R)}$.

In what follows, we replace π by $\pi_{\exp(-\beta R)}$ for a given $\beta > 0$, keeping one more degree of freedom. Now, note that Theorem 1.5 gives:

$$\begin{aligned} &\rho(R) - R(\theta') \\ &\quad \leq \rho(r) - r(\theta') + \frac{\lambda}{2N} \overline{\mathbb{P}} \left[\int_{\Theta} (l_\theta - l_{\theta'})^2 \, d\rho(\theta) \right] + \frac{\mathcal{K}\left(\rho, \pi_{\exp(-\beta R)}\right) + \log \frac{1}{\varepsilon}}{\lambda} \end{aligned}$$

and note that the upper bound is no longer empirical (observable to the statistician).

The aim of the next subsection is to upper bound $\mathcal{K}\left(\rho, \pi_{\exp(-\beta R)}\right)$ by an empirical bound in a general setting.

2.2. Empirical bound of the localized complexity.

Definition 2.1. Let us put, for any $(a, \lambda) \in]0,1] \times \mathbb{R}_+^*$ and $(\theta, \theta') \in \Theta^2$,

$$v_{a, \frac{\lambda}{N}}(\theta, \theta') = \frac{2N}{\lambda} \left\{ \frac{\lambda}{N} \sum_{i=1}^{N} \Phi_{\frac{\lambda}{N}} \left[(\ell_\theta - \ell_{\theta'})(Z_i) \wedge \frac{aN}{\lambda} \right] - \Big[r(\theta) - r(\theta') \Big] \right\}.$$

Theorem 2.1. *Let us choose a distribution $\pi \in \mathcal{M}_+^1(\Theta)$. For any $\varepsilon > 0$, for any $(a, \gamma, \beta) \in]0,1] \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ such that $\beta < \gamma$, with $\mathbb{P}$-probability at least $1 - \varepsilon$, for any $\rho \in \mathcal{M}_+^1(\Theta)$,*

$$\mathcal{K}\left(\rho, \pi_{\exp(-\beta R)}\right) \leq \mathcal{BK}_{a,\beta,\gamma}(\rho, \pi) + \frac{\beta}{\gamma - \beta} \log \frac{1}{\varepsilon}$$

where

$$\begin{aligned} \mathcal{BK}_{a,\beta,\gamma}(\rho, \pi) = \left(1 - \frac{\beta}{\gamma}\right)^{-1} \Bigg\{ &\mathcal{K}\left(\rho, \pi_{\exp(-\beta r)}\right) \\ &+ \log \int_{\Theta} \pi_{\exp(-\beta r)}(d\theta') \exp\left[\int_{\Theta} \rho(d\theta) \left(\frac{\beta\gamma}{2N} v_{a,\frac{\gamma}{N}}(\theta, \theta') + \beta \Delta_{\frac{\gamma}{a}}(\theta, \theta') \right) \right] \Bigg\}. \end{aligned}$$

The proof is given in the section dedicated to proofs, more precisely in subsection 5.3 page 16. Note that the localized entropy term is controlled by its empirical counterpart together with a variance term.

Before combining this result with Theorem 1.5, we give the analogous result for the non-integrated case, which proof is also given in subsection 5.3.

Theorem 2.2. *Let us choose a distribution $\pi \in \mathcal{M}_+^1(\Theta)$ and a randomized estimator ρ. For any $\varepsilon > 0$ and $\eta > 0$, for any $(a,\gamma,\beta) \in]0,1] \times \mathbb{R}_+^* \times \mathbb{R}_+^*$ such that $\beta < \gamma$, with $\mathbb{P}\rho$-probability at least $1-\varepsilon$,*

$$\log\left[\frac{d\rho}{d\pi_{\exp[-\beta R]}}(\tilde{\theta})\right] \leq \mathcal{D}_{a,\beta,\gamma}(\rho,\pi)(\tilde{\theta}) + \frac{\beta}{\gamma-\beta}\log\frac{1}{\varepsilon}$$

where

$$\mathcal{D}_{a,\beta,\gamma}(\rho,\pi)(\tilde{\theta}) = \left(1-\frac{\beta}{\gamma}\right)^{-1}\left\{\log\left[\frac{d\rho}{d\pi_{\exp[-\beta r]}}(\tilde{\theta})\right]\right.$$
$$\left. + \log\int_{\Theta}\pi_{\exp(-\beta r)}(d\theta')\exp\left[\frac{\beta\gamma}{2N}v_{a,\frac{\gamma}{N}}(\tilde{\theta},\theta') + \beta\Delta_{\frac{\gamma}{a}}(\tilde{\theta},\theta')\right]\right\}.$$

2.3. **Localized PAC-Bayesian theorems.**

Definition 2.2. From now on, we will deal with model selection. We assume that we have a family of submodels of Θ: $(\Theta_i, i \in I)$ where I is finite or countable. We also choose a probability measure $\mu \in \mathcal{M}_+^1(I)$, and assume that we have a prior distribution $\pi^i \in \mathcal{M}_+^1(\Theta_i)$ for every i.

We choose

$$\pi = \sum_{i\in I}\mu(i)\pi^i_{\exp(-\beta_i R)}$$

and apply Theorem 1.3 that we combine with Theorem 2.2 by a union bound argument, to obtain the following result.

Theorem 2.3. *Let us assume that we have randomized estimators $(\rho_i)_{i\in I}$ such that $\rho_i(\Theta_i) = 1$, for any $\varepsilon > 0$, for any $(a,\beta,\beta',\gamma,\gamma',\lambda) \in]0,1] \times (\mathbb{R}_+^*)^5$ such that $\beta < \gamma$ and $\beta' < \gamma'$, with $\mathbb{P}\bigotimes_{i\in I}\rho_i$-probability at least $1-\varepsilon$ over the sample $(Z_n)_{n=1,\dots,N}$ and the parameters $(\tilde{\theta}_i)_{i\in I}$, for any $(i,i') \in I^2$ we have:*

$$R_{\frac{\lambda}{a}}\big(\tilde{\theta}_i,\tilde{\theta}_{i'}\big) \leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{r\big(\tilde{\theta}_i\big) - r\big(\tilde{\theta}_{i'}\big) + \frac{\lambda}{2N}v_{a,\frac{\lambda}{N}}\big(\tilde{\theta}_i,\tilde{\theta}_{i'}\big)\right.$$
$$+\frac{1}{\lambda}\left[\mathcal{D}_{a,\beta,\gamma}(\rho,\pi^i)\big(\tilde{\theta}_i\big) + \mathcal{D}_{a,\beta',\gamma'}(\rho,\pi^{i'})\big(\tilde{\theta}_{i'}\big)\right.$$
$$\left.\left. + \left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right]\right\}.$$

In the same way, we can give an integrated variant, using Theorem 1.5 and Theorem 2.1.

Theorem 2.4. *For any $\varepsilon > 0$, for any $(a,\beta,\beta',\gamma,\gamma',\lambda) \in]0,1] \times (\mathbb{R}_+^*)^5$ such that $\beta < \gamma$ and $\beta' < \gamma'$, with $\mathbb{P}$-probability at least $1-\varepsilon$, for any $(i,i') \in I^2$ and $(\rho,\rho') \in \mathcal{M}_+^1(\Theta_i) \times \mathcal{M}_+^1(\Theta_{i'})$,*

$$\int_{\Theta^2} d(\rho\otimes\rho')(\theta,\theta')R_{\frac{\lambda}{a}}(\theta,\theta')$$
$$\leq \Phi^{-1}_{\frac{\lambda}{N}}\left\{\rho(r) - \rho'(r) + \frac{\lambda}{2N}\int_{\Theta^2} d(\rho\otimes\rho')(\theta,\theta')\,v_{a,\frac{\lambda}{N}}(\theta,\theta')\right.$$
$$\left. + \frac{\mathcal{BK}_{a,\beta,\gamma}(\rho,\pi^i) + \mathcal{BK}_{a,\beta',\gamma'}(\rho',\pi^{i'}) + \left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\mu(i)\mu(i')}}{\lambda}\right\}.$$

2.4. **Choice of the parameters.** In this subsection, we explain how to choose the parameters λ, β, β', γ and γ' in Theorems 2.3 and 2.4. In some really simple situations (parametric model with strong assumptions on $\mathbb{P}$), this choice can be made on the basis of theoretical considerations, however, in many realistic situations, such hypothesis cannot be made and we would like to optimize the upper bound in the Theorems with respect to the parameters. This would lead to data-dependant values for the parameters, and this is not allowed by Theorems 2.4 and 2.3. Catoni [8] proposes to make a union bound on a grid of values of the parameters, thus allowing optimization with respect to these parameters. We apply this idea to Theorem 2.4, and obtain the following result.

Theorem 2.5. *Let us choose a measure $\nu \in \mathcal{M}_+^1(\Theta)$ that is supported by a finite or countable set of points, $supp(\nu)$. Let us assume that we have randomized estimators $(\rho_{i,\beta})_{i\in I,\beta\in supp(\nu)}$ such that $\rho_{i,\beta}(\Theta_i)=1$. For any $\varepsilon>0$ and $a\in]0,1]$, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon$ over the sample $(Z_n)_{n=1,\ldots,N}$ and the parameters $(\tilde{\theta}_{i,\beta})_{i\in I,\beta\in supp(\nu)}$, for any $(i,i')\in I^2$ and $(\beta,\beta')\in supp(\nu)^2$ we have:*

$$
\begin{aligned}
&R_{\frac{\lambda}{a}}\big(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\big)\\
&\qquad \leq B\Big((i,\beta),(i',\beta')\Big) = \inf_{\substack{\lambda\in]0,+\infty[\\ \gamma\in]\beta,+\infty[\\ \gamma'\in]\beta',+\infty[}} \Phi_{\frac{\lambda}{N}}^{-1}\Bigg\{ r\big(\tilde{\theta}_{i,\beta}\big) - r\big(\tilde{\theta}_{i',\beta'}\big)\\
&+\frac{\lambda}{2N}v_{a,\frac{\lambda}{N}}\big(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\big) + \frac{1}{\lambda}\Bigg[\mathcal{D}_{a,\beta,\gamma}(\rho_{i,\beta},\pi^i)\big(\tilde{\theta}_{i,\beta}\big) + \mathcal{D}_{a,\beta',\gamma'}(\rho_{i,\beta'},\pi^{i'})\big(\tilde{\theta}_{i',\beta'}\big)\\
&\qquad + \left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\nu(\lambda)\nu(\gamma)\nu(\beta)\nu(\gamma')\nu(\beta')\mu(i)\mu(i')}\Bigg]\Bigg\}.
\end{aligned}
$$

2.5. **Introduction of the complexity function.** It is convenient to remark that we can dissociate the optimization with respect to the different parameters in Theorem 2.5 thanks to the introduction of an appropriate complexity function. The model selection algorithm we propose in the next subsection takes advantage of this decomposition.

Definition 2.3. Let us choose some real constants $\zeta>1$, $a\in]0,1]$ and $\varepsilon>0$. We assume that some randomized estimators $(\rho_{i,\beta})_{i\in I,\beta\in supp(\nu)}$ have been chosen and that we have drawn $\tilde{\theta}_{i,\beta}$ for every $i\in I$ and $\beta\in supp(\nu)$. We define, for any $i\in I$,

$$
\begin{aligned}
\mathcal{C}\big(i,\beta\big) = \inf_{\gamma\in[\zeta\beta,+\infty[}\Bigg\{&\mathcal{D}_{a,\beta,\gamma}(\rho_{i,\beta},\pi^i)\big(\tilde{\theta}_{i,\beta}\big)\\
&+\left(\frac{\beta}{\gamma-\beta}+\frac{1}{\zeta-1}+1\right)\log\frac{3}{\varepsilon\mu(i)\nu(\beta)\nu(\gamma)}\Bigg\}.
\end{aligned}
$$

We have the following result.

Theorem 2.6. *For any $(i,i',\beta,\beta')\in I^2\times supp(\nu)^2$,*

$$
\begin{aligned}
B\Big((i,\beta),(i',\beta')\Big) \leq \inf_{\lambda>0}\Phi_{\frac{\lambda}{N}}^{-1}\Bigg\{& r\big(\tilde{\theta}_{i,\beta}\big) - r\big(\tilde{\theta}_{i',\beta'}\big)\\
&+\frac{\lambda}{2N}v_{a,\frac{\lambda}{N}}\big(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\big) + \frac{\mathcal{C}\big(\tilde{\theta}_{i,\beta}\big)+\mathcal{C}\big(\tilde{\theta}_{i',\beta'}\big)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}\Bigg\}.
\end{aligned}
$$

Note, as a consequence of the concavity of $\Phi_{\frac{\lambda}{N}}^{-1}$, that this implies

Corollary 2.7.

$$B\Big((i,\beta),(i',\beta')\Big)+B\Big((i',\beta'),(i,\beta)\Big) \leq 2\inf_{\lambda>0}\Phi^{-1}_{\frac{\lambda}{N}}\left\{\frac{\lambda}{2N}\frac{v_{a,\frac{\lambda}{N}}\left(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\right)+v_{a,\frac{\lambda}{N}}\left(\tilde{\theta}_{i',\beta'},\tilde{\theta}_{i,\beta}\right)}{2} + \frac{\mathcal{C}\left(i,\beta\right)+\mathcal{C}\left(i',\beta'\right)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}\right\}.$$

Corollary 2.7 shows that the symmetric part of B has an upper bound which contains only variance and complexity terms.

3. Application: model selection

In this section, we propose a general algorithm to select among a family of posteriors - and so to perform model selection as a particular case. This algorithm was introduced by Catoni [8] in the case of classification. We first give the general form of the estimator. We then give an empirical bound on its risk. The last subsection is devoted to a theoretical bound under suitable hypothesis.

3.1. **Selection algorithm.**

We introduce the following definition for the sake of simplicity.

Definition 3.1. Let us put:

$$\mathcal{P}=\left\{t^1,...,t^M\right\}=\left\{(i,\beta)\in I\times supp(\nu)\right\},$$

where $M=|I|\times|supp(\nu)|$ and the indexation of the t_i's is such that

$$\mathcal{C}(t^1)\leq ...\leq \mathcal{C}(t^M).$$

Now, remark that there is no reason for the bound B defined in Theorem 2.5 to be sub-additive. So let us define a sub-additive version of B.

Definition 3.2. We put, for any $(t,t')\in\mathcal{P}^2$:

$$\tilde{B}(t,t')=\inf_{\substack{h\geq 1\\ (t_0,...,t_h)\in\mathcal{P}^{h+1}\\ t_0=t,\, t_h=t'}}\sum_{k=1}^{h}B(t_{k-1},t_k).$$

Definition 3.3. For any $k\in\{1,...,M\}$ we put:

$$s(k)=\inf\left\{j\in\{1,...,M\},\quad \tilde{B}(t^k,t^j)>0\right\}.$$

We are now ready to give the definition of our estimator.

Definition 3.4. We take as an estimator $\tilde{\theta}_{\hat{t}}$ where $\hat{t}=t^{\hat{k}}$ and

$$\hat{k}=\min\left(\arg\max s\right).$$

3.2. **Empirical bound on the risk of the selected estimator.**

Theorem 3.1. *Let us put $\hat{s}=s(\hat{k})$. For any $\varepsilon>0$, with $\mathcal{P}\bigotimes_{t\in\mathcal{P}}\rho_t$-probability at least $1-\varepsilon$,*

$$R\left(\tilde{\theta}_{\hat{t}}\right)\leq R\left(\tilde{\theta}_{t^j}\right)+\begin{cases}0, & 1\leq j<\hat{s},\\ \tilde{B}(t^{s(j)},t^j) & \hat{s}\leq j<\hat{k},\\ \tilde{B}(\hat{t},t^{\hat{s}})+\tilde{B}(t^{\hat{s}},t^j), & j\in(\arg\max s)\\ \tilde{B}(\hat{t},t^j), & \textit{otherwise.}\end{cases}$$

Thus, adding only non negative terms to the bound,

$$R\left(\tilde{\theta}_{\hat{t}}\right) \leq R\left(\tilde{\theta}_{t^j}\right) + \begin{cases} 0, & 1 \leq j < \hat{s}, \\ B(t^{s(j)}, t^j) + B(t^j, t^{s(j)}) & \hat{s} \leq j < \hat{k}, \\ B(t^j, t^{\hat{s}}) + B(t^{\hat{s}}, t^j) & \\ \qquad + B(\hat{t}, t^{\hat{s}}) + B(t^{\hat{s}}, \hat{t}\,) & j \in (\arg\max s), \\ B(t^j, \hat{t}\,) + B(\hat{t}, t^j), & \text{otherwise}. \end{cases}$$

For a proof, we refer the reader to Catoni [8] where this Theorem is proved in the case of classification, the proof can be reproduced here without any modification.

Theorem 3.1 shows that, according to Corollary 2.7 (page 12), $R(\tilde{\theta}_{\hat{t}}) - R(\tilde{\theta}_{t^j})$ can be bounded by variance and complexity terms relative to posterior distributions with a complexity not greater than $\mathcal{C}(t^j)$, and an empirical loss in any case not much larger than the one of $\tilde{\theta}_{t^j}$.

3.3. **Theoretical bound.** In this subsection, we choose $\rho_{i,\beta}$ as $\pi^i_{\exp(-\beta r)}$ restricted to a (random) neighborhood of $\hat{\theta}_i$. More formally, for any $p \geq 0$, let us put

$$\Theta_{i,p} = \left\{\theta \in \Theta_i, \quad r(\theta) - \inf_{\Theta_i} r \leq p\right\}$$

and for any $q \in]0,1]$ let us put

$$p_{i,\beta}(q) = \inf\left\{p > 0, \quad \pi^i_{\exp(-\beta r)}(\Theta_{i,p}) \geq q\right\}.$$

Then let us choose q once and for all and let us choose $\rho_{i,\beta}$ so that

$$\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}(\theta) = \frac{\mathbb{1}_{\Theta_{i,p_{i,\beta}(q)}}(\theta)}{\pi^i_{\exp(-\beta r)}\left(\Theta_{i,p_{i,\beta}(q)}\right)}.$$

Moreover, we assume that $0 \leq l_\theta(z) \leq C$ for any $\theta \in \Theta$ and $z \in \mathcal{Z}$, and we fix $a = 1$. In this case, note that for any $\lambda \leq N/(2C)$ we have:

$$v_{1,\frac{\lambda}{N}}(\theta,\theta') \leq \overline{\mathbb{P}}\left[(l_\theta - l_{\theta'})^2\right].$$

For the sake of simplicity we introduce the following definition.

Definition 3.5. Let us put, for any $(\theta,\theta') \in \Theta^2$:

$$v(\theta,\theta') = \overline{\mathbb{P}}\left[(l_\theta - l_{\theta'})^2\right]$$

and

$$V(\theta,\theta') = \mathbb{P}\left[v(\theta,\theta')\right].$$

To obtain the following result we take ν as the uniform measure on the grid

$$supp(\nu) = \left\{2^0, 2^1, ..., 2^{\lfloor \frac{\log N}{\log 2} \rfloor}\right\}.$$

Theorem 3.2. *Let us put, for any $i \in I$,*

$$\overline{\theta}_i = \arg\min_{\theta\in\Theta_i} R(\theta)$$

and

$$\overline{\theta} = \arg\min_{\theta\in\Theta} R(\theta).$$

Let us assume that Mammen and Tsybakov's margin assumption is satisfied, in other words let there exist $(\kappa, c) \in [1,+\infty[\times \mathbb{R}_+^*$ such that

$$\forall\theta \in \Theta, \quad \left[V(\theta,\overline{\theta})\right]^\kappa \leq c\left[R(\theta) - R(\overline{\theta})\right].$$

Let moreover every sub-model $\Theta_i, i \in I$ satisfy the following dimension assumption:

$$\sup_{\xi \in \mathbb{R}} \left\{ \xi \left[\pi^i_{\exp(-\xi R)}(R) - R\left(\overline{\theta}_i\right) \right] \right\} \leq d_i$$

for a given sequence $(d_i)_{i\in I} \in (\mathbb{R}_+)^I$. Then there is a constant $\mathcal{C} = \mathcal{C}(\kappa, c, C)$ such that, with $\mathbb{P} \bigotimes_{i\in I, \beta \in supp(\nu)} \rho_{i,\beta}$-probability at least $1 - 4\varepsilon$,

$$R\big(\tilde{\theta}_{\hat{i}}\big) \leq \inf_{i\in I} \Bigg\{ R\big(\overline{\theta}_i\big) + \mathcal{C} \max \Bigg\{ \left(\frac{\left[R\left(\overline{\theta}_i\right) - R\left(\overline{\theta}\right)\right]^{\frac{1}{\kappa}} \left(d_i + \log \frac{1}{q} + \log \frac{1+\log_2 N}{\varepsilon \mu(i)} \right)}{N} \right)^{\frac{1}{2}}, \left(\frac{d_i + \log \frac{1}{q} + \log \frac{1+\log_2 N}{\varepsilon\mu(i)}}{N} \right)^{\frac{\kappa}{2\kappa - 1}} \Bigg\} \Bigg\}.$$

For a proof, see subsection 5.4 page 16. Let us now make some remarks.

Remark 3.1 (Choice of the parameter q). The better choice for q is obviously $q = 1$. In this case, our estimator is drawn randomly from the distribution,

$$\rho_{i,\beta} = \pi^i_{\exp(-\beta r)},$$

and the term $\log(1/q)$ vanishes.

However, practitioners worried about the idea to choose randomly in the whole space an estimator can use a smaller value of q ensuring that, in any model i and for any β,

$$r\left(\tilde{\theta}_{i,\beta}\right) \leq \inf_{\Theta_i} r + p_{i,\beta}(q),$$

so $\tilde{\theta}_{i,\beta}$ is drawn in a neighborhood of the minimizer of the empirical risk.

Remark 3.2 (Margin assumption). The so-called margin assumption

$$\left[V(\theta, \overline{\theta})\right]^{\kappa} \leq c\left[R(\theta) - R(\overline{\theta})\right]$$

was first introduced by Mammen and Tsybakov in the context of classification [15, 20]. It has however been studied in the context of general regression by Lecue in his PhD Thesis [14]. The terminology comes from classification, where a similar assumption can be described in terms of margin. In the general case however, there is no margin involved, but rather a distance $V(\theta, \theta')^{1/2}$ on the parameter space, which serves to describe the shape of the function R in the neighborhood of its minimum value $R(\overline{\theta})$.

Remark 3.3 (Dimension assumption). In many cases, the assumption

$$\sup_{\xi \in \mathbb{R}} \left\{ \xi \left[\pi^i_{\exp(-\xi R)}(R) - R\left(\overline{\theta}_i\right) \right] \right\} \leq d_i$$

is just equivalent to the fact that every Θ_i has a finite dimension proportionnal to d_i.

4. Conclusion

In this paper we studied a quite general regression problem. We proposed randomized estimators, that can we drawn in small neighborhoods of empirical minimizers. We proved that these estimators reach the minimax rate of convergence under Mammen and Tsybakov's margin assumption.

We would like also to point out that the techniques used here can be applied in a more general context. In particular, Catoni [8] studied the transductive classification setting, where for a given $k \in \mathbb{N}$, we observe the objects $X_1, \cdots, X_{(k+1)N}$ and the labels $Y_1, \cdots, Y_N$, and we want to predict the kN missing labels $Y_{N+1}, \cdots, Y_{(k+1)N}$. In this context, a deviation result equivalent to Lemma 1.2 (page 6) can

be proved, and from this result we can obtain a theorem similar to Theorem 3.1 (page 12). We refer the reader to our PhD thesis [1] for more details (the transductive setting is introduced page 54 and the deviation result is Lemma 3.1 page 56).

5. Proofs

5.1. **Proof of Lemma 1.1.** For the sake of completeness, we reproduce here the proof of Lemma 1.1 given in Catoni [6].

Proof of Lemma 1.1. Let us assume that h is upper-bounded on the support of n. Let us remark that m is absolutely continuous with respect to n if and only if it is absolutely continuous with respect to $n_{\exp(h)}$. If it is the case, then

$$\mathcal{K}\left(m, n_{\exp(h)}\right) = m\left\{\log\left(\frac{dm}{dn}\right) - h\right\} + \log n(\exp\circ h) \\ = \mathcal{K}(m,n) - m(h) + \log n(\exp\circ h).$$

The left-hand side of this equation is nonnegative and cancels only for $m = n_{\exp(h)}$. Note that it remains valid when m is not absolutely continuous with respect to n and just says in this case that $+\infty = +\infty$. We therefore obtain

$$0 = \inf_{m\in\mathcal{M}_+^1(E)} \left[\mathcal{K}(m,n) - m(h)\right] + \log n(\exp\circ h).$$

This proves the second part of lemma 1.1. For the first part, we do not assume any longer that h is upper bounded on the support of n. We can write

$$\log n(\exp\circ h) = \sup_{B\in\mathbb{R}} \log n[\exp\circ(h\wedge B)] = \sup_{B\in\mathbb{R}} \sup_{m\in\mathcal{M}_+^1(E)} \left[m\left(h\wedge B\right) - \mathcal{K}(m,n)\right] \\ = \sup_{m\in\mathcal{M}_+^1(E)} \sup_{B\in\mathbb{R}} \left[m\left(h\wedge B\right) - \mathcal{K}(m,n)\right] \\ = \sup_{m\in\mathcal{M}_+^1(E)} \left\{\sup_{B\in\mathbb{R}} \left[m\left(h\wedge B\right)\right] - \mathcal{K}(m,n)\right\} = \sup_{m\in\mathcal{M}_+^1(E)} \left[m(h) - \mathcal{K}(m,n)\right].$$

□

5.2. **Proof of Theorem 1.5.**

Proof of Theorem 1.5. The beginning of this proof follows exactly the proof of Theorem 1.3 (page 7) until Equation 1.2. Now, let us apply (to Equation 1.2) Lemma 1.1 with $(E,\mathcal{E}) = (\Theta^2, \mathcal{T}^{\otimes 2})$ to obtain:

$$\mathbb{P}\exp\left\{\sup_{m\in\mathcal{M}_+^1(\Theta^2)}\left[\int_{\Theta^2}\left\{\lambda\Phi_{\frac{\lambda}{N}}\left[R_{\frac{\lambda}{a}}(\theta,\theta')\right]\right.\right.\right. \\ \left.\left.\left. - \frac{\lambda}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\left(\ell_\theta - \ell_{\theta'}\right)(Z_i)\wedge\frac{aN}{\lambda}\right]\right\}dm(\theta,\theta') - \mathcal{K}(m,\pi\otimes\pi')\right]\right\} = 1.$$

Consequently

$$\mathbb{P}\exp\left\{\sup_{(\rho,\rho')\in[\mathcal{M}_+^1(\Theta)]^2}\left[\int_{\Theta^2}\left\{\lambda\Phi_{\frac{\lambda}{N}}\left[R_{\frac{\lambda}{a}}(\theta,\theta')\right]\right.\right.\right. \\ \left.\left.\left. - \frac{\lambda}{N}\sum_{i=1}^{N}\Phi_{\frac{\lambda}{N}}\left[\left(\ell_\theta - \ell_{\theta'}\right)(Z_i)\wedge\frac{aN}{\lambda}\right]\right\}d(\rho\otimes\rho')(\theta,\theta') - \mathcal{K}(\rho,\pi) - \mathcal{K}(\rho',\pi')\right]\right\} = 1.$$

This ends the proof. □

5.3. **Proof of Theorems 2.1 and 2.2.**

Proof of Theorem 2.1. First, notice that:

$$\mathcal{K}\left(\rho,\pi_{\exp(-\beta R)}\right)=\beta\left[\rho(R)-\pi_{\exp(-\beta R)}(R)\right]+\mathcal{K}\left(\rho,\pi\right)-\mathcal{K}\left(\pi_{\exp(-\beta R)},\pi\right).$$

Let us apply Theorem 1.5 with $\pi=\pi'=\rho'=\pi_{\exp(-\beta R)}$ to obtain with probability at least $1-\varepsilon$, for any $\rho\in\mathcal{M}_{+}^{1}(\Theta)$,

$$\mathcal{K}\left(\rho,\pi_{\exp(-\beta R)}\right)\leq\beta\Bigg[\rho(r)-\pi_{\exp(-\beta R)}(r)$$
$$+\frac{\gamma}{2N}\int_{\Theta^2}v_{a,\frac{\gamma}{N}}(\theta,\theta')d\big(\rho\otimes\pi_{\exp(-\beta R)}\big)(\theta,\theta')+\frac{\log\frac{1}{\varepsilon}+\mathcal{K}\left(\rho,\pi_{\exp(-\beta R)}\right)}{\gamma}$$
$$+\int_{\Theta^2}\Delta_{\frac{\lambda}{a}}(\theta,\theta')d\big(\rho\otimes\pi_{\exp(-\beta R)}\big)(\theta,\theta')\Bigg]+\mathcal{K}\left(\rho,\pi\right)-\mathcal{K}\left(\pi_{\exp(-\beta R)},\pi\right).$$

Replacing in the right-hand side of this inequality $\pi_{\exp(-\beta R)}$ with a supremum over all possible distributions leads to the announced result. □

Proof of Theorem 2.2. We have, for any θ:

$$\log\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\theta)=\beta\left[R(\theta)-\pi_{\exp(-\beta R)}(R)\right]+\log\frac{d\rho}{d\pi}(\theta)-\mathcal{K}\left(\pi_{\exp(-\beta R)},\pi\right).$$

Let us apply Theorem 1.3 with $\pi=\pi'=\rho'=\pi_{\exp(-\beta R)}$ and a general ρ to obtain with $\mathbb{P}\rho$-probability at least $1-\varepsilon$ over θ,

$$\log\frac{d\rho}{d\pi_{\exp(-\beta R)}}(\theta)\leq\beta\Bigg[r(\theta)-\pi_{\exp(-\beta R)}(r)$$
$$+\frac{\gamma}{2N}\int_{\Theta}v_{a,\frac{\gamma}{N}}(\theta,\theta')d\pi_{\exp(-\beta R)}(\theta')+\frac{\log\frac{1}{\varepsilon}+\mathcal{K}\left(\rho,\pi_{\exp(-\beta R)}\right)}{\gamma}$$
$$+\int_{\Theta}\Delta_{\frac{\lambda}{a}}(\theta,\theta')d\pi_{\exp(-\beta R)}(\theta')\Bigg]+\log\frac{d\rho}{d\pi}(\theta)-\mathcal{K}\left(\pi_{\exp(-\beta R)},\pi\right).$$

The end of the proof is the same as in the case of Theorem 2.1. □

5.4. **Proof of Theorem 3.2.** We begin by a set of preliminary lemmas and definitions.

Definition 5.1. For the sake of simplicity, we will write:

$$r'(\theta,\theta')=r(\theta)-r(\theta')$$

and

$$R'(\theta,\theta')=R(\theta)-R(\theta')$$

for any $(\theta,\theta')\in\Theta^2$.

Definition 5.2. We introduce the margin function:

$$\varphi:\mathbb{R}_{+}^{*}\to\mathbb{R}$$
$$x\mapsto\sup_{\theta\in\Theta}\Big[V(\theta,\overline{\theta})-xR'(\theta,\overline{\theta})\Big].$$

Lemma 5.1 (Mammen and Tsybakov [15, 20])**.** *Mammen's and Tsybakov margin assumtion:*

$$\exists(\kappa,c)\in[1,+\infty[\times\mathbb{R}_{+}^{*},\forall\theta\in\Theta,\quad V(\theta,\overline{\theta})^{\kappa}\leq cR'(\theta,\overline{\theta})$$

implies:

$$\forall x > 0, \quad \varphi(x) \leq \left(1 - \frac{1}{\kappa}\right) (\kappa c x)^{-\frac{1}{\kappa-1}}$$

for $\kappa > 1$ *and* $\varphi(c) \leq 0$ *for* $\kappa = 1$.

Definition 5.3. We define the modified Bernstein function:

$$g : \mathbb{R} \to \mathbb{R}$$
$$x \mapsto \begin{cases} \frac{2[\exp(x)-1-x]}{x^2} & \text{if} \quad x \neq 0, \\ 1 \quad \text{if} \quad x = 0. \end{cases}$$

The function g is a variant of Bernstein's function, used in Bernstein's inequality (see Bernstein [4]). Here, we prove a variant of this inequality.

Lemma 5.2 (Variant of Bernstein's inequality)**.** *We have, for any* $\lambda > 0$ *and any* $(\theta, \theta') \in \Theta^2$*:*

$$\mathbb{P} \exp\left[\lambda R'(\theta,\theta') - \lambda r'(\theta,\theta') - \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(\theta,\theta')\right] \leq 1, \tag{5.1}$$

and the reverse inequality

$$\mathbb{P} \exp\left[\lambda r'(\theta,\theta') - \lambda R'(\theta,\theta') - \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(\theta,\theta')\right] \leq 1. \tag{5.2}$$

We also have a similar inequality for variances:

$$\mathbb{P} \exp\left[\frac{N}{4C^2} v(\theta,\theta') - \frac{N}{2C^2} V(\theta,\theta')\right] \leq 1. \tag{5.3}$$

Proof. We have:

$$\mathbb{P} \exp\left[\lambda R'(\theta,\theta') - \lambda r'(\theta,\theta')\right]$$
$$= \exp\left\{\sum_{i=1}^{N} \log \mathbb{P} \exp\left[-\frac{\lambda}{N}\left(l_\theta - l_{\theta'}\right)(Z_i)\right] + \lambda R'(\theta,\theta')\right\}.$$

Now, note that for any $b > 0$, for any $x \in [-b, b]$ we have:

$$\exp(-x) = 1 - x + \frac{x^2}{2} g(-x) \leq 1 - x + \frac{x^2}{2} g(b),$$

so that

$$\log \mathbb{P} \exp\left[-\frac{\lambda}{N}\left(l_\theta - l_{\theta'}\right)(Z_i)\right] \leq -\lambda R'(\theta,\theta') + \frac{\lambda^2}{2N} g\left(\frac{2C\lambda}{N}\right) V(\theta,\theta').$$

It shows that

$$\mathbb{P} \exp\left[\lambda R'(\theta,\theta') - \lambda r'(\theta,\theta')\right] \leq \exp\left[\frac{\lambda^2}{2N} g\left(\frac{2C\lambda}{N}\right) V(\theta,\theta')\right].$$

The proof of the reverse inequality follows the same scheme. For Inequality (5.3) note that, using the same scheme, we obtain:

$$\mathbb{P} \exp\left\{\lambda v(\theta,\theta') - \lambda V(\theta,\theta') - \frac{\lambda^2}{2N} g\left(\frac{4\lambda C^2}{N}\right) \mathbb{P}\left[\left(l_\theta - l_{\theta'}\right)^4 (Z)\right]\right\} \leq 1.$$

This implies that

$$\mathbb{P} \exp\left[\lambda v(\theta,\theta') - \lambda V(\theta,\theta') - \frac{\lambda^2 2C^2}{N} g\left(\frac{4\lambda C^2}{N}\right) V(\theta,\theta')\right] \leq 1.$$

The choice $\lambda = N/4C^2$ and the remark that $g(1) \leq 2$ (actually $g(1) \simeq 1.4$) leads to Inequality (5.3). $\square$

Definition 5.4. For the sake of shortness, we put:

$$\delta_N(i,q,\varepsilon,\kappa) = \max\Bigg\{ \left(\frac{\left[R\left(\overline{\theta}_i\right) - R\left(\overline{\theta}\right)\right]^{\frac{1}{\kappa}} \left(d_i + \log\frac{1}{q} + \log\frac{1+\log_2 N}{\varepsilon\mu(i)}\right)}{N} \right)^{\frac{1}{2}}, \left(\frac{d_i + \log\frac{1}{q} + \log\frac{1+\log_2 N}{\varepsilon\mu(i)}}{N} \right)^{\frac{\kappa}{2\kappa-1}} \Bigg\}.$$

Now let us give a brief overview of what follows. Lemma 5.3 proves that for some β, $\tilde{\theta}_{i,\beta}$ achieves the expected rate of convergence in model Θ_i: $\delta_N(i,q,\varepsilon,\kappa)$. As we then want to use Theorem 3.1 to compare our estimator $\tilde{\theta}_{\hat{i}}$ to every possible $\tilde{\theta}_{i,\beta}$, we will have to control the various parts of the empirical bound $B(.,.)$ by theoretical terms. So we give two more lemmas: Lemma 5.4 controls the empirical variance term $v(.,.)$ by the theoretical variance term $V(.,.)$ while Lemma 5.5 provides a control for the empirical complexity term $\mathcal{C}(i,\beta)$. Given these three results we will be able to prove Theorem 3.2. Let us start with

Lemma 5.3. *Under the assumptions of Theorem 3.2, there is a constant $\mathcal{C}' = \mathcal{C}'(\kappa,c,C)$ such that, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon$, for any $i \in I$, there is a $\beta = \beta^*(i) \in supp(\nu)$ such that*

$$R'\left(\tilde{\theta}_{i,\beta}, \overline{\theta}_i\right) \leq \mathcal{C}'\delta_N(i,q,\varepsilon,\kappa).$$

Proof. We have, by Inequality (5.1) in Lemma 5.2:

$$\begin{aligned} 1 &\geq \pi^i_{\exp(-\beta R)}\mathbb{P}\exp\left[\lambda R'(.,\overline{\theta}_i) - \lambda r'(.,\overline{\theta}_i) - \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(.,\overline{\theta}_i)\right] \\ &\geq \mathbb{P}\rho_{i,\beta}\exp\Bigg[\lambda R'(.,\overline{\theta}_i) - \lambda r'(.,\overline{\theta}_i) \\ &\qquad - \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(.,\overline{\theta}_i) - \log\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta R)}}(.)\Bigg]. \end{aligned}$$

Thus

$$\begin{aligned} \mu(i)\nu(\beta) \geq \mathbb{P}\rho_{i,\beta}\exp\Bigg[&\lambda R'(.,\overline{\theta}_i) - \lambda r'(.,\overline{\theta}_i) \\ &- \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(.,\overline{\theta}_i) - \log\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta R)}}(.) + \log(\mu(i)\nu(\beta))\Bigg]. \end{aligned}$$

So, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon/2$, for any $i \in I$ and $\beta \in supp(\nu)$,

$$\begin{aligned} \lambda R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \lambda r'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) &+ \frac{\lambda^2}{2N} g\left(\frac{2\lambda C}{N}\right) V(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \\ &+ \log\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta R)}}(\tilde{\theta}_{i,\beta}) + \log\frac{2}{\varepsilon\mu(i)\nu(\beta)}. \end{aligned} \tag{5.4}$$

Note that, using Definition 5.2, for any $x > 0$,

$$V(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq 2\left[V(\tilde{\theta}_{i,\beta},\overline{\theta}) + V(\overline{\theta},\overline{\theta}_i)\right] \leq 2\left[xR'(\tilde{\theta}_{i,\beta},\overline{\theta}) + xR'(\overline{\theta}_i,\overline{\theta}) + 2\varphi(x)\right].$$

Therefore Inequality (5.4) becomes:

$$\left[\lambda - \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right)\right] R'(\tilde{\theta}_{i,\beta},\overline{\theta}) \leq \left[\lambda + \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right)\right] R'(\overline{\theta}_i,\overline{\theta})$$
$$+ \lambda r'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) + \frac{2\varphi(x)\lambda^2}{N} g\left(\frac{2\lambda C}{N}\right) + \beta\left[R\left(\tilde{\theta}_{i,\beta}\right) - \pi^i_{\exp(-\beta R)} R\right]$$
$$+ \log \frac{d\rho_{i,\beta}}{d\pi^i}(\tilde{\theta}_{i,\beta}) - \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) + \log \frac{2}{\varepsilon\mu(i)\nu(\beta)},$$

leading to

$$\left[\lambda - \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) - \beta\right] R'(\tilde{\theta}_{i,\beta},\overline{\theta}) \leq \left[\lambda + \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) - \beta\right] R'(\overline{\theta}_i,\overline{\theta})$$
$$+ \frac{2\varphi(x)\lambda^2}{N} g\left(\frac{2\lambda C}{N}\right) - \beta\pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i) + \log \frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}(\tilde{\theta}_{i,\beta})$$
$$- \log \pi^i \exp\left[-\lambda r'(.,\overline{\theta}_i)\right] - \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) + \log \frac{2}{\varepsilon\mu(i)\nu(\beta)}$$

and

$$\left[\lambda - \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) - \beta\right] R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \frac{2x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) R'(\overline{\theta}_i,\overline{\theta}) \tag{5.5}$$
$$+ \frac{2\varphi(x)\lambda^2}{N} g\left(\frac{2\lambda C}{N}\right) - \beta\pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i) + \log \frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}(\tilde{\theta}_{i,\beta})$$
$$- \log \pi^i \exp\left[-\lambda r'(.,\overline{\theta}_i)\right] - \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) + \log \frac{2}{\varepsilon\mu(i)\nu(\beta)}.$$

We can then use Inequality (5.2) (in Lemma 5.2, page 17) to obtain, with probability at least $1-\varepsilon/2$, for any $i \in I$ and $\beta \in supp(\nu)$,

$$- \log \pi^i \exp\left[-\lambda r'(.,\overline{\theta}_i)\right] \leq \lambda\pi^i_{\exp(-\beta R)} r'(.,\overline{\theta}_i) + \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) \tag{5.6}$$
$$\leq \lambda\pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i) + \frac{\lambda^2}{2N} g\left(\frac{2C\lambda}{N}\right) \pi^i_{\exp(-\beta R)} V(.,\overline{\theta}_i)$$
$$+ \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) - \log \frac{\varepsilon\mu(i)\nu(\beta)}{2}$$
$$\leq \left[\lambda + \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right)\right] \pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i) + \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) R'(\overline{\theta}_i,\overline{\theta})$$
$$+ \frac{2\varphi(x)\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) + \mathcal{K}\left(\pi^i_{\exp(-\beta R)}, \pi^i\right) + \log \frac{2}{\varepsilon\mu(i)\nu(\beta)}.$$

Combining Inequalities (5.5) and (5.6) we have, with probability at least $1-\varepsilon$, for any i and β:

$$\left[\lambda - \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) - \beta\right] R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \frac{4x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) R'(\overline{\theta}_i,\overline{\theta}) \tag{5.7}$$
$$+ \frac{4\varphi(x)\lambda^2}{N} g\left(\frac{2\lambda C}{N}\right) + \left[\lambda + \frac{x\lambda^2}{N} g\left(\frac{2C\lambda}{N}\right) - \beta\right] \pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i)$$
$$+ \log \frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}(\tilde{\theta}_{i,\beta}) + 2\log \frac{2}{\varepsilon\mu(i)\nu(\beta)}.$$

In order to make explicit the terms in Inequality 5.7, let us remind the definition of $\rho_{i,\beta}$ in Theorem 3.2 (page 13) and remark that

$$\log \frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}(\tilde{\theta}_{i,\beta}) \leq \log \frac{1}{q}.$$

Let us also recall the dimension hypothesis in Theorem 3.2, implying that

$$\pi^i_{\exp(-\beta R)} R'(.,\overline{\theta}_i) \leq \frac{d_i}{\beta}.$$

Let us finally choose $\lambda = 2\beta$, Inequality 5.7 becomes:

$$\left[\beta - \frac{16x\beta^2}{N} g\left(\frac{4C\beta}{N}\right)\right] R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \frac{16x\beta^2}{N} g\left(\frac{4C\beta}{N}\right) R'(\overline{\theta}_i,\overline{\theta}) + \frac{16\varphi(x)\beta^2}{N} g\left(\frac{4\beta C}{N}\right) + \left[\beta + \frac{4x\beta^2}{N} g\left(\frac{4C\beta}{N}\right)\right] d_i + \log\frac{1}{q} + 2\log\frac{2}{\varepsilon\mu(i)\nu(\beta)}. \tag{5.8}$$

Finally, Lemma 5.1 together with the margin assumption in Theorem 3.2 ensures that

$$\varphi(x) \leq \left(1 - \frac{1}{\kappa}\right)(\kappa c x)^{\frac{-1}{\kappa-1}}$$

if $\kappa > 1$ and $\varphi(c) \leq 0$ if $\kappa = 1$. Let us first deal with the case $\kappa = 1$. Inequality (5.8) becomes, taking $x = c$,

$$R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \left[\frac{1}{2} - \frac{4c\beta}{N} g\left(\frac{4C\beta}{N}\right)\right]^{-1} \left\{\frac{16c\beta}{N} g\left(\frac{4C\beta}{N}\right) R'(\overline{\theta}_i,\overline{\theta}) + \left[1 + \frac{4c\beta}{N} g\left(\frac{4C\beta}{N}\right)\right]\frac{d_i}{\beta} + \frac{1}{\beta}\log\frac{1}{q} + \frac{2}{\beta}\log\frac{2}{\varepsilon\mu(i)\nu(\beta)}\right\}. \tag{5.9}$$

In the right-hand side of Inequality 5.9, the numerator is optimal for β of the order of

$$\sqrt{\frac{N\left(d_i + \log\frac{1}{q} + \log\frac{2}{\varepsilon\mu(i)\nu(\beta)}\right)}{R'(\overline{\theta}_i,\overline{\theta})}}$$

but in order to keep the denominator away from zero, the maximal order of magnitude for β is N, so let us take β of the order of

$$\min\left\{\sqrt{\frac{N\left(d_i + \log\frac{1}{q} + \log\frac{2}{\varepsilon\mu(i)\nu(\beta)}\right)}{R'(\overline{\theta}_i,\overline{\theta})}}, N\right\}.$$

This choice leads to:

$$R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \mathcal{C}'' \max\left\{\left(\frac{\left[R\left(\overline{\theta}_i,\overline{\theta}\right)\right]\left(d_i + \log\frac{1}{q} + \log\frac{1+\log_2 N}{\varepsilon\mu(i)}\right)}{N}\right)^{\frac{1}{2}}, \left(\frac{d_i + \log\frac{1}{q} + \log\frac{1+\log_2 N}{\varepsilon\mu(i)}}{N}\right)\right\} = \mathcal{C}''\delta_N(i,q,\varepsilon,1) \tag{5.10}$$

for some $\mathcal{C}'' = \mathcal{C}''(c,C)$. In the case where $\kappa > 1$, Inequality (5.8) becomes:

$$R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i) \leq \left[\frac{1}{2} - \frac{4x\beta}{N} g\left(\frac{4C\beta}{N}\right)\right]^{-1} \left\{\frac{16x\beta}{N} g\left(\frac{4C\beta}{N}\right) R'(\overline{\theta}_i,\overline{\theta}) + \left(1 - \frac{1}{\kappa}\right)\frac{16\beta(\kappa c x)^{-\frac{1}{\kappa-1}}}{N} g\left(\frac{4\beta C}{N}\right) + \left[1 + \frac{4x\beta}{N} g\left(\frac{4C\beta}{N}\right)\right]\frac{d_i}{\beta} + \frac{1}{\beta}\log\frac{1}{q} + \frac{2}{\beta}\log\frac{2}{\varepsilon\mu(i)\nu(\beta)}\right\}. \tag{5.11}$$

Now, we choose x or the order of

$$\min\left\{\left[R'\left(\overline{\theta}_i,\overline{\theta}\right)\right]^{-\frac{\kappa-1}{\kappa}},\frac{N}{\beta}\right\}$$

in Inequality (5.11) (the case $x=\left[R'\left(\overline{\theta}_i,\overline{\theta}\right)\right]^{-\frac{\kappa-1}{\kappa}}$ minimizes the numerator while the fact that $x=\mathcal{O}(N/\beta)$ ensures that the denominator does not get too close to zero). Now, let us consider both cases for x, and first $x=\left[R'\left(\overline{\theta}_i,\overline{\theta}\right)\right]^{-\frac{\kappa-1}{\kappa}}$. In this case, let us choose β of the order of

$$\min\left\{\sqrt{\frac{N\left(d_i+\log\frac{1}{q}+\log\frac{2}{\varepsilon\mu(i)\nu(\beta)}\right)}{\left[R'(\overline{\theta}_i,\overline{\theta})\right]^{\frac{1}{\kappa}}}},N\right\}.$$

This leads to a bound of the order of

$$\max\left\{\left(\frac{\left[R\left(\overline{\theta}_i,\overline{\theta}\right)\right]^{\frac{1}{\kappa}}\left(d_i+\log\frac{1}{q}+\log\frac{1+\log_2 N}{\varepsilon\mu(i)}\right)}{N}\right)^{\frac{1}{2}},\right.$$
$$\left.\frac{d_i+\log\frac{1}{q}+\log\frac{1+\log_2 N}{\varepsilon\mu(i)}}{N}\right\}\leq\delta_N(i,q,\varepsilon,\kappa).$$

In the other case, x is of the order of N/β and

$$\left[R'\left(\overline{\theta}_i,\overline{\theta}\right)\right]^{-\frac{\kappa-1}{\kappa}}\geq\frac{N}{\beta},$$

implying that

$$R'\left(\overline{\theta}_i,\overline{\theta}\right)\leq\left(\frac{\beta}{N}\right)^{\frac{\kappa}{\kappa-1}}.$$

We have to choose β in order to optimize the numerator, in this case the optimal order of magnitude is

$$\left[\left(d_i+\log\frac{1}{q}+\log\frac{1+\log_2 N}{\varepsilon\mu(i)}\right)^{\kappa-1}N\right]^{\frac{1}{2\kappa-1}}$$

and leads to a bound of the order of

$$\left(\frac{d_i+\log\frac{1}{q}+\log\frac{1+\log_2 N}{\varepsilon\mu(i)}}{N}\right)^{\frac{\kappa}{2\kappa-1}}\leq\delta_N(i,q,\varepsilon,\kappa).$$

So we have proved that, in the case $\kappa>1$, for some $\mathcal{C}'''=\mathcal{C}'''(\kappa,c,C)$,

$$R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i)\leq\mathcal{C}'''\delta_N(i,q,\varepsilon,\kappa). \tag{5.12}$$

We put:

$$\mathcal{C}'(\kappa,c,C)=\begin{cases}\mathcal{C}''(c,C) & \text{if}\quad\kappa=1\\ \\ \mathcal{C}'''(\kappa,c,C) & \text{if}\quad\kappa>1\end{cases}$$

and remark that Inequalities (5.10) and (5.12) end the proof. □

Lemma 5.4. *Under the assumptions of Theorem 3.2, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon$, for any $(i,i')\in I^2$, for any $(\beta,\gamma,\beta',\gamma')\in supp(\nu)^4$:*

$$v\left(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\right)\leq 2V\left(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\right)+\frac{4C^2}{N}\left[\mathcal{D}_{1,\beta,\gamma}\left(\rho_{i,\beta},\pi^i\right)\left(\tilde{\theta}_{i,\beta}\right)\right.$$

$$+\mathcal{D}_{1,\beta',\gamma'}\left(\rho_{i',\beta'},\pi^{i'}\right)\left(\tilde{\theta}_{i',\beta'}\right)+\left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\Bigg].$$

Proof. According to Inequality (5.3) (Lemma 5.2 page 17),

$$\mathbb{P}\exp\left[\frac{N}{4C^2}v(\theta,\theta')-\frac{N}{4C^2}V'(\theta,\theta')\right]\leq 1.$$

Let us integrate in (θ,θ') with respect to the distribution $\pi^i_{\exp(-\beta R)}\otimes\pi^{i'}_{\exp(-\beta' R)}$ and sum over all i, i', β and β' to obtain, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon/3$, for any $(i,i')\in I^2$, for any $(\beta,\beta')\in supp(\nu)^2$:

$$\begin{aligned}v\left(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\right)&\leq 2V'\left(\tilde{\theta}_{i,\beta},\tilde{\theta}_{i',\beta'}\right)\\&\quad+\frac{4C^2}{N}\left\{\log\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta R)}}\left(\tilde{\theta}_{i,\beta}\right)+\log\frac{d\rho_{i',\beta'}}{d\pi^{i'}_{\exp(-\beta' R)}}\left(\tilde{\theta}_{i',\beta'}\right)+\log\frac{3}{\varepsilon}\right\}.\end{aligned}$$

To conclude the proof, there remains to combine this result with Theorem 2.2 page 10, using a union bound argument. □

Lemma 5.5. *Under the assumptions of Theorem 3.2, there is a constant $K=K(\kappa,c,C)$ such that, with $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-\varepsilon$, for any $i\in I$, there is $\gamma\in supp(\nu)$ such that, for $\beta=\beta^*(i)$,*

$$\mathcal{D}_{1,\beta,\gamma}(\rho_{i,\beta},\pi^i)\left(\tilde{\theta}_{i,\beta}\right)\leq\mathcal{C}(i,\beta)\leq K\delta_N(i,q,\varepsilon,\kappa)\beta.$$

Proof. We have

$$\begin{aligned}&\mathcal{D}_{1,\beta,\gamma}(\rho_{i,\beta},\pi^i)\left(\tilde{\theta}_{i,\beta}\right)\\&=\left(1-\frac{\beta}{\gamma}\right)^{-1}\left\{\log\frac{d\rho_{i,\beta}}{d\pi^i_{\exp(-\beta r)}}\left(\tilde{\theta}_{i,\beta}\right)+\log\pi^i_{\exp(-\beta r)}\exp\left[\frac{\beta\gamma}{2N}v\left(.,\tilde{\theta}_{i,\beta}\right)\right]\right\}\\&\qquad\leq\left(1-\frac{\beta}{\gamma}\right)^{-1}\left\{\log\frac{1}{q}+\log\pi^i\exp\left[\frac{\beta\gamma}{2N}v\left(.,\tilde{\theta}_{i,\beta}\right)-\beta r'(.,\overline{\theta})\right]\right.\\&\qquad\qquad\left.-\log\pi^i\exp\left[-\beta r'(.,\overline{\theta})\right]\right\}.\end{aligned}$$

Let us now apply Lemma 5.2 and the now usual integration technique to obtain the following inequalities, with probability at least $1-4\varepsilon/5$:

$$\begin{aligned}-\log\pi^i\exp\bigl[-\beta r'(.,\overline{\theta})\bigr]&=-\sup_{\rho\in\mathcal{M}^1_+(\Theta_i)}\left[-\beta\rho r'(.,\overline{\theta})-\mathcal{K}(\rho,\pi^i)\right]\\&\leq-\sup_{\rho\in\mathcal{M}^1_+(\Theta_i)}\left[-\beta\rho R'(.,\overline{\theta})+\frac{\beta^2}{2N}g\left(\frac{2\beta C}{N}\right)V(.,\overline{\theta})+\log\frac{5}{\varepsilon}-\mathcal{K}(\rho,\pi^i)\right]\\&\leq-\log\pi^i\exp\left(-\beta R'(.,\overline{\theta})+\frac{\beta^2}{2N}g\left(\frac{2\beta C}{N}\right)V(.,\overline{\theta})\right)+\log\frac{5}{\varepsilon}.\end{aligned}$$

Moreover

$$\begin{aligned}&\log\pi^i\exp\left[\frac{\beta\gamma}{2N}v\left(.,\tilde{\theta}_{i,\beta}\right)-\beta r'(.,\overline{\theta})\right]\\&\qquad\leq\log\pi^i\exp\left\{\frac{\beta\gamma}{N}V(.,\tilde{\theta}_{i,\beta})\beta R'(.,\overline{\theta})+\frac{\beta^2}{2N}g\left(\frac{2\beta C}{N}\right)V(.,\overline{\theta})\right\}\\&\qquad\quad+\frac{\beta\gamma 4C^2}{N^2}\mathcal{D}_{1,\beta,\gamma}(\rho_{i,\beta},\pi^i)\left(\tilde{\theta}_{i,\beta}\right)+\left[1+\frac{4\beta\gamma C^2}{N^2}+\frac{4\gamma C^2\beta^2}{N(\gamma-\beta)}\right]\log\frac{5}{\varepsilon},\end{aligned}$$

so that

$$\left[1-\frac{\beta}{\gamma}-\frac{\beta\gamma 4C^2}{N^2}\right]\mathcal{D}_{1,\beta,\gamma}(\rho_{i,\beta},\pi^i)\left(\tilde{\theta}_{i,\beta}\right)$$
$$\leq \log\frac{1}{q}+\log\pi^i_{\exp(-\beta R)}\exp\left\{\left[\frac{\beta\gamma}{N}+\frac{\beta^2}{N}g\left(\frac{2\beta C}{N}\right)\right]V(.,\overline{\theta})\right\}$$
$$+\left[2+\frac{4\beta\gamma C^2}{N^2}+\frac{4\gamma C^2\beta^2}{N(\gamma-\beta)}\right]\log\frac{5}{\varepsilon}$$
$$\leq \log\frac{1}{q}+\log\pi^i_{\exp(-\beta R)}\exp\left\{x\left[\frac{\beta\gamma}{N}+\frac{\beta^2}{N}g\left(\frac{2\beta C}{N}\right)\right]R'(.,\overline{\theta}_i)\right\}$$
$$+\left[\frac{2\beta\gamma}{N}+\frac{\beta^2}{N}g\left(\frac{2\beta C}{N}\right)\right]\left[xR'(\overline{\theta}_i,\overline{\theta})+\varphi(x)\right]+x\frac{\beta\gamma}{N}R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i)$$
$$+\left[2+\frac{4\beta\gamma C^2}{N^2}+\frac{4\gamma C^2\beta^2}{N(\gamma-\beta)}\right]\log\frac{5}{\varepsilon}.$$

We then apply Lemma 5.3 to obtain with probability at least $1-\varepsilon/5$

$$R'(\tilde{\theta}_{i,\beta},\overline{\theta}_i)\leq \mathcal{C}'\delta_N(i,q,\varepsilon/5,\kappa).$$

Moreover we can choose $\gamma=2\beta$, and remember that the choice $\beta=\beta^*(i)$ leads to $\beta<N$, so

$$\left[\frac{1}{2}-\frac{\beta^2 8C^2}{N^2}\right]\mathcal{D}_{1,\beta,2\beta}(\rho_{i,\beta},\pi^i)\big(\tilde{\theta}_{i,\beta}\big)\tag{5.13}$$
$$\leq \log\frac{1}{q}+\log\pi^i_{\exp(-\beta R)}\exp\left\{\frac{x\beta^2}{N}\left[2+g\left(2C\right)\right]R'(.,\overline{\theta}_i)\right\}$$
$$+\frac{\beta^2}{N}\left[2+g\left(2C\right)\right]\left[xR'(\overline{\theta}_i,\overline{\theta})+\varphi(x)\right]+\frac{2x\beta^2}{N}\mathcal{C}'\delta_N(i,q,\varepsilon/5,\kappa)$$
$$+\left[2+\frac{12\beta^2C^2}{N^2}\right]\log\frac{5}{\varepsilon}.$$

Now, let us compute:

$$\log\pi^i_{\exp(-\beta R)}\exp\left\{\frac{x\beta^2}{N}\left[2+g\left(2C\right)\right]R'(.,\overline{\theta}_i)\right\}$$
$$\leq\frac{\beta^2}{N}\left[2+g(2C)\right]\int_0^x\pi^i_{\exp\left\{-\beta\left[1-\frac{\delta\beta}{N}(2+g(2C))\right]\right\}}R'(.,\overline{\theta}_i)d\delta$$
$$\leq\frac{\frac{\beta^2}{N}\left[2+g(2C)\right]}{\beta\left\{1-\frac{x\beta}{N}\left[2+g(2C)\right]\right\}}x\pi^i_{\exp\left\{-\beta\left[1-\frac{x\beta}{N}(2+g(2C))\right]\right\}}R'(.,\overline{\theta}_i)$$
$$\leq\frac{xd_i\beta}{N}\frac{2+g(2C)}{1-\frac{x\beta}{N}\left[2+g(2C)\right]}$$

by the dimension assumption, and so for any x smaller than N/β, Inequality 5.13 becomes

$$\left[\frac{1}{2}-\frac{\beta^2 8C^2}{N^2}\right]\mathcal{D}_{1,\beta,2\beta}(\rho_{i,\beta},\pi^i)\big(\tilde{\theta}_{i,\beta}\big)\tag{5.14}$$
$$\leq 2\beta\mathcal{C}'\delta_N(i,q,\varepsilon/5,\kappa)+\beta\left\{\frac{1}{\beta}\log\frac{1}{q}+\frac{d_i}{\beta}\frac{2+g(2C)}{1-\frac{x\beta}{N}\left[2+g(2C)\right]}\right.$$

$$+ \frac{\beta}{N}\left[2 + g\left(2C\right)\right]\left[xR'(\overline{\theta}_i,\overline{\theta}) + \varphi(x)\right] + \frac{2+12g(2C)}{\beta}\log\frac{5}{\varepsilon}\Bigg\}.$$

The optimization of the right-hand side of Inequality (5.14) with respect to x and β leads to the same discussion as for the optimization of the right-hand side of Inequality (5.8) (page 20) in the proof of Lemma 5.3 (and a choice of x satisfying $x < N/\beta$). □

We are now able to proceed to the

proof of Theorem 3.2. With $\mathbb{P}\bigotimes_{i\in I,\beta\in supp(\nu)}\rho_{i,\beta}$-probability at least $1-4\varepsilon$ the inequalities stated in Theorem 3.1 and in Lemmas 5.3, 5.4 and 5.5 are simultaneously satisfied. In this case, let us choose $i \in I$, $\beta = \beta^*(i)$ and j such that $t^j = (i,\beta)$. We have:

$$R'\big(\tilde{\theta}_{\hat{t}},\tilde{\theta}_{(i,\beta)}\big) \leq \begin{cases} 0, & 1 \leq j < \hat{s} \quad \text{(case 1)},\\ B(t^{s(j)},t^j) + B(t^j,t^{s(j)}) & \hat{s} \leq j < \hat{k} \quad \text{(case 2)},\\ B(t^j,t^{\hat{s}}) + B(t^{\hat{s}},t^j) & \\ \quad + B(\hat{t},t^{\hat{s}}) + B(t^{\hat{s}},\hat{t}) & j \in (\arg\max s) \quad \text{(case 3)},\\ B(t^j,\hat{t}) + B(\hat{t},t^j), & \text{otherwise (case 4)}.\end{cases}$$

Let us examine successively the four cases (1, 2, 4 and 3, this last case being the most difficult).

Case 1: if $1 \leq j < \hat{s}$, then

$$R'\big(\tilde{\theta}_{\hat{t}},\tilde{\theta}_{(i,\beta)}\big) \leq 0,$$

and so, by the result of Lemma 5.3 (page 18),

$$R'\big(\tilde{\theta}_{\hat{t}},\overline{\theta}_i\big) \leq \mathcal{C}'\delta_N(i,q,\varepsilon,\kappa).$$

Case 2: the idea in all the remaining cases (2, 4 and 3) is that we have to give a control of $R'\big(\tilde{\theta}_{\hat{t}},\tilde{\theta}_{(i,\beta)}\big)$, controlled by the empirical bound $B(.,.)$, in terms of theoretical quantities only. In case 2, $\hat{s} \leq j < \hat{k}$, then for any $\lambda \in supp(\nu)$,

$$\begin{aligned} R'\big(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{(i,\beta)}\big) &\leq B(t^{s(j)},t^j) + B(t^j,t^{s(j)}) \\ &\leq \frac{\lambda}{2N}v(t^{s(j)},t^j) + \frac{\mathcal{C}(t^{s(j)}) + \mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ &\leq \frac{\lambda}{N}V(t^{s(j)},t^j) + \frac{\mathcal{C}(t^j) + \mathcal{C}(t^{s(j)}) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ + \quad \frac{4C^2\lambda}{N^2}&\left[\mathcal{C}(t^j) \quad + \quad \mathcal{C}(t^{s(j)}) \quad + \quad \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right] \quad . \end{aligned}$$

As we have, by definition of the function $s(.)$, the inequality $\mathcal{C}(t^{s(j)}) \leq \mathcal{C}(t^j)$,

$$\begin{aligned} R'\big(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{(i,\beta)}\big) \leq \frac{\lambda}{N}V(t^{s(j)},t^j) &+ \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ &+ \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right], \end{aligned}$$

and so

$$\begin{aligned} R'\big(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{t^j}\big) &\leq \frac{2\lambda}{N}\left[xR'\big(\tilde{\theta}_{t^{s(j)}},\overline{\theta}\big) + xR'\big(\tilde{\theta}_{t^j},\overline{\theta}\big) + \varphi(x)\right] \\ &+ \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \end{aligned} \tag{5.15}$$

$$+ \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right].$$

Thus

$$\begin{aligned}\left[1 - \frac{2\lambda x}{N}\right] R'\bigl(\tilde{\theta}_{t^{s(j)}}, \tilde{\theta}_{t^j}\bigr) \leq \frac{2\lambda}{N}\left[2xR'(\tilde{\theta}_{t^j}, \overline{\theta}_i) + 2xR'(\overline{\theta}_i, \overline{\theta}) + \varphi(x)\right] \\ + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ + \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\right)\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right].\end{aligned}$$

Let us apply Lemma 5.5 page 22 to upper bound $\mathcal{C}(t^j)$, Lemma 5.3 page 18 to upper bound $R'(\tilde{\theta}_{t^j}, \overline{\theta}_i)$ and Lemma 5.1 page 16 to upper bound $\varphi(x)$. Let us put moreover $\lambda = \gamma = \gamma' = 2\beta = 2\beta'$ and remember that $\beta < N$. We obtain, for any x such that $x < N/\beta$,

$$\begin{aligned}\left[1 - \frac{4\beta x}{N}\right] R'\bigl(\tilde{\theta}_{t^{s(j)}}, \tilde{\theta}_{t^j}\bigr) \leq \frac{4\beta}{N}\left[2xR'\bigl(\overline{\theta}_i, \overline{\theta}\bigr) + \left(1 - \frac{1}{\kappa}\right)(\kappa c x)^{\frac{-1}{\kappa-1}}\right] \\ + [K(1+32C^2) + 8\mathcal{C}']\delta_N(i,q,\varepsilon,\kappa) + \frac{1}{2\beta} + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}2\beta \\ + \frac{48C^2}{N}\log\frac{3}{\varepsilon\mu(i)\mu(i')}.\end{aligned}$$

Let us replace x and β by the values given in the discussion for the optimization of the right-hand side of Inequality (5.8) (page 20) in the proof of Lemma 5.3 (and a choice of x satisfying $x < N/\beta$) to obtain the existence of a constant $\mathcal{D}' = \mathcal{D}'(\kappa, c, C)$ such that

$$R'\bigl(\tilde{\theta}_{t^{s(j)}}, \tilde{\theta}_{t^j}\bigr) \leq \mathcal{D}'\delta_N(i,q,\varepsilon,\kappa).$$

We then deduce from this result and from Lemma 5.3 that

$$R'\bigl(\tilde{\theta}_{\hat{t}}, \overline{\theta}_i\bigr) \leq R'\bigl(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{t^j}\bigr) + R'\bigl(\tilde{\theta}_{t^j}, \overline{\theta}_i\bigr) \leq (\mathcal{D}' + \mathcal{C}')\,\delta_N(i,q,\varepsilon,\kappa).$$

Case 4: the proof follows roughly the same scheme than for case 2; if $j > \max(\arg\max s)$, note that $\mathcal{C}(t^j) \geq \mathcal{C}(\hat{t})$, therefore

$$\begin{aligned}R'\bigl(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{(i,\beta)}\bigr) &\leq B(\hat{t}, t^j) + B(t^j, \hat{t}) \\ &\leq \frac{\lambda}{2N}v(\hat{t}, t^j) + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ &\leq \frac{\lambda}{N}V(\hat{t}, t^j) + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ &\quad + \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right. \\ &\leq \frac{2\lambda}{N}\left[xR'(\tilde{\theta}_{\hat{t}}, \overline{\theta}) + xR'(\tilde{\theta}_{t^j}, \overline{\theta}) + \varphi(x)\right] + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ &\quad + \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right].\end{aligned}$$

Thus

$$\left[1 - \frac{2\lambda x}{N}\right] R'\bigl(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{(i,\beta)}\bigr) \leq \frac{4\lambda x}{N}\left[R'\bigl(\tilde{\theta}_{(i,\beta)}, \overline{\theta}_i\bigr) + R'\left(\overline{\theta}_i, \overline{\theta}\right)\right]$$

$$+\frac{2\lambda\varphi(x)}{N}+\frac{2\mathcal{C}(t^j)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}$$
$$+\frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j)+\left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right].$$

Let us apply Lemma 5.5 page 22 to upper bound $\mathcal{C}(t^j)$, Lemma 5.3 page 18 to upper bound $R'(\tilde{\theta}_{t^j},\overline{\theta}_i)$ and Lemma 5.1 page 16 to upper bound $\varphi(x)$. Let us put moreover $\lambda=\gamma=\gamma'=2\beta=2\beta'$ and remember that $\beta<N$. We obtain, for any x such that $x<N/\beta$,

$$\left[1-\frac{4\beta x}{N}\right]R'\big(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{(i,\beta)}\big)\leq\frac{4\beta}{N}\left[2xR'(\overline{\theta}_i,\overline{\theta})+\left(1-\frac{1}{\kappa}\right)(\kappa c x)^{\frac{-1}{\kappa-1}}\right]$$
$$+[K(1+32C^2)+8\mathcal{C}']\delta_N(i,q,\varepsilon,\kappa)+\frac{1}{2\beta}+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}2\beta$$
$$+\frac{48C^2}{N}\log\frac{3}{\varepsilon\mu(i)\mu(i')}.$$

Choosing x exactly in the same way as in the previous cases and replacing $\beta=\beta^*(i)$ with its value, we obtain the existence of $\mathcal{D}''=\mathcal{D}''(\kappa,c,C)$ such that

$$R'\big(\tilde{\theta}_{\hat{t}},\tilde{\theta}_{(i,\beta)}\big)\leq\mathcal{D}''\delta_N(i,q,\varepsilon,\kappa)$$

and so

$$R\big(\tilde{\theta}_{\hat{t}},\overline{\theta}_i\big)\leq(\mathcal{C}'+\mathcal{D}'')\,\delta_N(i,q,\varepsilon,\kappa).$$

Case 3: if $j\in(\arg\max s)$, remember that $\hat{s}=s(\hat{t})=s(j)$, so that

$$R'\big(\tilde{\theta}_{\hat{t}},\tilde{\theta}_{t^j}\big)\leq\Big[B(t^j,t^{s(j)})+B(t^{s(j)},t^j)\Big]+\Big[B(\hat{t},t^{\hat{s}})+B(t^{\hat{s}},\hat{t})\Big].\tag{5.16}$$

We are going to upper bound separately $B(t^j,t^{s(j)})+B(t^{s(j)},t^j)$ and $B(\hat{t},t^{\hat{s}})+B(t^{\hat{s}},\hat{t})$. Let us first deal with the term $B(t^j,t^{s(j)})+B(t^{s(j)},t^j)$:

$$\Big[B(t^j,t^{s(j)})+B(t^{s(j)},t^j)\Big]\leq\frac{\lambda}{2N}v(t^{s(j)},t^j)+\frac{2\mathcal{C}(t^j)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}\tag{5.17}$$
$$\leq\frac{\lambda}{N}V(t^{s(j)},t^j)+\frac{2\mathcal{C}(t^j)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}$$
$$+\frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j)+\left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right.$$
$$\leq\frac{2\lambda}{N}\left[xR'(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{t^j})+2xR'(\tilde{\theta}_{t^j},\overline{\theta}_i)+2xR'(\overline{\theta}_i,\overline{\theta})+\varphi(x)\right]$$
$$+\frac{2\mathcal{C}(t^j)+\frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda}$$
$$+\frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j)+\left(1+\frac{\beta}{\gamma-\beta}+\frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right].$$

Let us notice that

$$R'(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{t^j})\leq B(t^{s(j)},t^j)$$

and remember that, by definition, $B(t^j,t^{s(j)})\geq 0$. This shows that

$$R'(\tilde{\theta}_{t^{s(j)}},\tilde{\theta}_{t^j})\leq\Big[B(t^j,t^{s(j)})+B(t^{s(j)},t^j)\Big].$$

Once again, let us apply Lemma 5.5 to upper bound $\mathcal{C}(t^j)$, Lemma 5.3 to upper bound $R'(\tilde{\theta}_{t^j},\overline{\theta}_i)$ and Lemma 5.1 to upper bound $\varphi(x)$. Let us put moreover $\lambda=$

$\gamma = \gamma' = 2\beta = 2\beta'$. Inequality 5.17 becomes:

$$\left(1 - \frac{4\beta x}{N}\right)\left[B(t^j, t^{s(j)}) + B(t^{s(j)}, t^j)\right] \leq \frac{4\beta}{N}\left[2xR'(\overline{\theta}_i, \overline{\theta}) + \left(1 - \frac{1}{\kappa}\right)(\kappa c x)^{\frac{-1}{\kappa-1}}\right] \\ + [K(1+32C^2) + 8\mathcal{C}']\delta_N(i,q,\varepsilon,\kappa) + \frac{1}{2\beta} + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}2\beta \\ + \frac{48C^2}{N}\log\frac{3}{\varepsilon\mu(i)\mu(i')}$$

and therefore

$$\left[B(t^j, t^{s(j)}) + B(t^{s(j)}, t^j)\right] \leq \mathcal{E}\delta_N(i,q,\varepsilon,\kappa).$$

There remains to upper bound $\left[B(\hat{t}, t^{\hat{s}}) + B(t^{\hat{s}}, \hat{t})\right]$. We will use to that purpose the fact that $\mathcal{C}(\hat{t}) \leq \mathcal{C}(t^j)$:

$$\left[B(\hat{t}, t^{\hat{s}}) + B(t^{\hat{s}}, \hat{t})\right] \\ \leq \frac{2\lambda}{N}\left[xR'(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{t^j}) + xR'(\tilde{\theta}_{t^{s(j)}}, \tilde{\theta}_{t^j}) + 2xR'(\tilde{\theta}_{t^j}, \overline{\theta}_i) + 2xR'(\overline{\theta}_i, \overline{\theta}) + \varphi(x)\right] \\ + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ + \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right].$$

Note that we have already proved that

$$R'(\tilde{\theta}_{t^{s(j)}}, \tilde{\theta}_{t^j}) \leq \left[B(t^j, t^{s(j)}) + B(t^{s(j)}, t^j)\right] \leq \mathcal{E}\delta_N(i,q,\varepsilon,\kappa).$$

Plugging all these results into Inequality (5.16), we obtain,

$$\left(1 - \frac{2\lambda x}{N}\right)R'(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{t^j}) \leq \mathcal{E}\delta_N(i,q,\varepsilon,\kappa) \\ + \frac{2\lambda}{N}\left[x\mathcal{E}\delta_N(i,q,\varepsilon,\kappa) + 2xR'(\tilde{\theta}_{t^j}, \overline{\theta}_i) + 2xR'(\overline{\theta}_i, \overline{\theta}) + \varphi(x)\right] \\ + \frac{2\mathcal{C}(t^j) + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}}{\lambda} \\ + \frac{4C^2\lambda}{N^2}\left[2\mathcal{C}(t^j) + \left(1 + \frac{\beta}{\gamma-\beta} + \frac{\beta'}{\gamma'-\beta'}\log\frac{3}{\varepsilon\mu(i)\mu(i')}\right)\right].$$

As usual, let us apply Lemma 5.5 to upper bound $\mathcal{C}(t^j)$, Lemma 5.3 to upper bound $R'(\tilde{\theta}_{t^j}, \overline{\theta}_i)$ and Lemma 5.1 to upper bound $\varphi(x)$. Let us put $\lambda = \gamma = \gamma' = 2\beta = 2\beta'$, to obtain

$$\left(1 - \frac{4\beta x}{N}\right)R'\left(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{t^j}\right) \leq \frac{4\beta}{N}\left[2xR'(\overline{\theta}_i, \overline{\theta}) + \left(1 - \frac{1}{\kappa}\right)(\kappa c x)^{\frac{-1}{\kappa-1}}\right] \\ + [K(1+32C^2) + 8\mathcal{C}' + 3\mathcal{E}]\delta_N(i,q,\varepsilon,\kappa) + \frac{1}{2\beta} + \frac{\zeta+1}{\zeta-1}\log\frac{3}{\varepsilon\nu(\lambda)}2\beta \\ + \frac{48C^2}{N}\log\frac{3}{\varepsilon\mu(i)\mu(i')},$$

and therefore

$$R'\left(\tilde{\theta}_{\hat{t}}, \tilde{\theta}_{t^j}\right) \leq \mathcal{E}'\delta_N(i,q,\varepsilon,\kappa).$$

This ends the proof. □

Appendix : bounding the effect of truncation

We will show here how to upper bound $R(\theta)-R(\theta')-R_\lambda(\theta,\theta')$ by some quantity $\Delta_\lambda(\theta,\theta')$ depending on an additional hypothesis on the data distribution.

Lemma 5.6. *Let us assume that we are in the i.i.d. case, where $P_1 = ... = P_N$ and that for some constants $(b,B) \in \mathbb{R}_+^2$*

$$\forall\theta \in \Theta, \quad P_1\left\{\exp\left[b\left|l_\theta(Z_1)\right|\right]\right\} \leq B.$$

Then, for any $(\theta,\theta') \in \Theta^2$,

$$R(\theta)-R(\theta')-R_\lambda(\theta,\theta') \leq \Delta_\lambda(\theta,\theta') = \frac{2B}{b}\exp\left(\frac{-bN}{2\lambda}\right).$$

Proof. From definitions,

$$\begin{aligned}
&R(\theta)-R(\theta')-R_\lambda(\theta,\theta')\\
&\quad = P_1\left\{l_\theta(Z_1)-l_{\theta'}(Z_1)-\left[l_\theta(Z_1)-l_{\theta'}(Z_1)\right]\wedge\frac{N}{\lambda}\right\}\\
&\quad = P_1\left[\left(l_\theta(Z_1)-l_{\theta'}(Z_1)-\frac{N}{\lambda}\right)_+\right],
\end{aligned}$$

where $(x)_+ = x\wedge 0$. So we can write

$$\begin{aligned}
&R(\theta)-R(\theta')-R_\lambda(\theta,\theta')\\
&\quad \leq \int_0^{+\infty} P_1\left[\left(l_\theta(Z_1)-l_{\theta'}(Z_1)-\frac{N}{\lambda}\right)_+ > t\right]dt\\
&\quad \leq \int_0^{+\infty} P_1\left[l_\theta(Z_1)-l_{\theta'}(Z_1)-\frac{N}{\lambda} > t\right]dt\\
&\quad \leq \int_0^{+\infty} P_1\left\{\exp\left[\frac{b}{2}\left(l_\theta(Z_1)-l_{\theta'}(Z_1)-\frac{N}{\lambda}-t\right)\right]\right\}dt\\
&\quad \leq \exp\left(\frac{-bN}{2\lambda}\right)B\int_0^{+\infty}\exp\left(-\frac{bt}{2}\right)dt,
\end{aligned}$$

leading to the result stated in the lemma. □

References

[1] Alquier, P. *Transductive and Inductive Adaptative Inference for Regression and Density Estimation.* PhD thesis, University Paris 6, 2006.

[2] Audibert, J.-Y. Aggregated estimators and empirical complexity for least square regression. *Annales de l'Institut Henri Poincarï£¡: Probability and Statistics 40*, 6 (2004), 685–736.

[3] Audibert, J.-Y. *PAC-Bayesian Statistical Learning Theory.* PhD thesis, University Paris 6, 2004.

[4] Bernstein, S. N. *Theory of Probability (in Russian).* Moscow, 1927.

[5] Boucheron, S., Bousquet, O., and Lugosi, G. Theory of classification: some recent advances. *ESAIM Probability and Statistics 9* (2005), 323–375.

[6] Catoni, O. A pac-bayesian approach to adaptative classification. *Preprint Laboratoire de Probabilitï£¡s et Modï£¡les Alï£¡atoires* (2003).

[7] Catoni, O. *Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001).* Springer, 2004.

[8] Catoni, O. *PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)*, vol. 56 of *Lecture Notes-Monograph Series.* IMS, 2007.

[9] Cervonenkis, A. J., and Vapnik, V. N. On the uniform convergence of relative frequencies of events to their probabilities. *Doklady Akademii Nauk USSR 181*, 4 (1968).

[10] Cortes, C., and Vapnik, V. N. Support vector networks. *Machine Learning 20* (1995), 273–297.

[11] Devroye, L., Gyï£¡rfi, L., and Lugosi, G. *A Probabilistic Theory of Pattern Recognition.* Springer, New-York, 1996.

[12] Friedman, J. H., Hastie, T., and Tibshirani, R. *The Elements of Statistical Learning.* Springer, New-York, 2001.

[13] Hoeffding, W. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association 18* (1963), 13–30.

[14] Lecue, G. *Aggregation Procedures: Optimality and Fast Rates.* PhD thesis, University Paris 6, 2007.

[15] Mammen, E., and Tsybakov, A. Smooth discrimination analysis in statistical learning. *The Annals of Statistics 27* (1999), 1808–1829.

[16] McAllester, D. A. Some pac-bayesian theorems. In *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (Madison, WI, 1998)* (1998), ACM, pp. 230–234.

[17] McAllester, D. A. Pac-bayesian model averaging. In *Proceedings of of the Twelth Annual Conference On Computational Learning Theory, Santa Cruz, California (Electronic)* (1999), ACM, New-York, pp. 164–170.

[18] Rissanen, J. Modeling by shortest data description. *Automatica 14* (1978), 465–471.

[19] Schapire, R., Freund, Y., Bartlett, P., and Lee, W. S. Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics 26* (1998), 1651–1686.

[20] Tsybakov, A. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics 32* (2004), 135–166.

[21] Vapnik, V. N. *The Nature of Statistical Learning Theory.* Springer-Verlag, 1998.

[22] Zhang, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *The Annals of Statistics 32* (2004), 56–85.

CREST, and, Laboratoire de Probabilités et Modèles Aléatoires (Université Paris 7), 175, rue du Chevaleret, 75252 Paris Cedex 05, France.

URL: http://alquier.ensae.net/

E-mail address: alquier@ensae.fr